

THE DETERMINACY OF FACTOR SCORE MATRICES WITH IMPLICATIONS FOR FIVE OTHER BASIC PROBLEMS OF COMMON-FACTOR THEORY¹

By LOUIS GUTTMAN

Israel Institute of Applied Social Research

I. Introduction. II. The Problem of Determinacy. III. Supplementary Problems

I. INTRODUCTION

1. *Problem.* Common-factor analysis—in the sense of Spearman, Thurstone and others, begins by considering the scores of a given population of N individuals on each of n quantitative variables from a given universe of content. Its goal is to express the observed scores as linear combinations of scores on common- and unique-factors.

It is typical of factor analysis procedures that they approach their goal only indirectly. If ξ is the observed score matrix, a *direct* analysis would consist in finding factor score matrices η and ζ such that

$$\xi = A\eta + \zeta \quad (1)$$

where A is some real matrix of common-factor loadings. Instead of a direct resolution into factor scores such as (1), indirect procedures are used, which pivot on the observed correlation matrix, say R . Real matrices A , L , and U are sought, where L is Gramian and U is diagonal, such that

$$R = ALA' + U^2. \quad (2)$$

As is well known, unless further restrictions are imposed, infinitely many solutions always exist for the right member of (2). There are different schools of thought as to the conditions to be imposed on U , especially on the rank of $R - U^2$; this is the well-known problem of 'communalities'. There are also different schools of thought as to what conditions should be imposed on A and L ; this is the problem of 'rotation of axes'. Irrespective of their differences, each school in practice seems to regard its main job as completed for a given ξ when A , L , and U have been computed for (2) to satisfy that school's particular criteria for communalities and rotations.

The purpose of the present paper is to explore the extent to which the indirect analysis of the scores in ξ via (2) is equivalent to a direct analysis of the type (1). Part of this problem has been studied previously—from a somewhat different point of view—by other writers, notably E. B. Wilson, Godfrey Thomson, and Walter Ledermann [see discussion and references in 15, pp. 371 f.]. In the present paper, we shall establish the existence of solutions to (1) given (2), and a set of necessary and sufficient conditions for the construction of all possible solutions. This will enable us to go further and examine in detail some basic questions concerning the scientific meaning of common-factor analysis.

2. *Sufficiency Versus Necessity: The Six Problems.* The reason why (2) has entered the picture of factor analysis is that it is a *necessary* condition for (1). If A , η , and ζ are given such that (1) is satisfied, if L and U^2 are the covariance matrices of the common- and unique-factor scores respectively, and if the unique-factor scores have zero correlations with each other and with all the common-factors, then (2) necessarily follows.

¹ This research was facilitated in part by an uncommitted grant-in-aid from the Behavioral Sciences Division of the Ford Foundation, and in part by a grant for methodological work from the Lucius N. Littauer Foundation.

The Determinacy of Factor Score Matrices

Necessity is not the same thing as sufficiency. Computations concerned only with (2) yield only matrices A , L , and U^2 . To reach the goal of the factor analysis requires one to ask further: does there exist a pair of score matrices, η and ξ , to satisfy (1) with the given A and with the covariances specified by the given L and U^2 ? If yes, how many such pairs of score matrices exist for the fixed A , L , and U^2 ? That is, to what extent is (2) *sufficient* for establishing a solution to (1)?

This existence problem turns out always to have an affirmative answer, but also to have *too many* affirmative answers, especially when $U^2 \neq 0$. Furthermore, for finite n , these alternative answers can differ widely among themselves. We shall establish conditions under which (when $N = \infty$) there will be essentially but one answer in the limit as $n \rightarrow \infty$.

The multiplicity of solutions for η and ξ when n is finite—even when A , L , and U^2 are fixed—has important implications for the psychological meaning of common-factor analysis, as well as for the computing procedures conventionally used. If we denote as ‘Problem I’ the determinacy of η and ξ , solving Problem I throws light on the five further basic problems listed below. These are discussed more fully in the last part of the paper.

Problem II. Communalities. The problem of the ‘unknown communalities’—or equivalently, of the choice of U^2 —turns out to be related to Problem I. We show a unique solution is definable under certain conditions as $n \rightarrow \infty$.

Problem III. Meaning of factors. It has been thought by many that the meaning of common-factor scores in η could be ascertained from A in (1), or factor variables in η could be *named* according to the observed variables in ξ which have high loadings on them. This may be illusory in the light of the answer to Problem I, for η itself can have widely different simultaneous solutions no matter what the structure of a fixed A may be.

Problem IV. ‘Inverted’ factor analysis. Merely considering the case where $N = \infty$, as for example when the variables in ξ are continuous, raises questions, which may not have been previously apparent, in respect of the proposal to ‘factor’ people rather than variables.

Problem V. ‘Second-order’ common factors. When L in (2) is not a diagonal matrix, some researchers have proposed ‘factoring’ it in turn like R . Solving Problem I shows that ‘second-order’ scores may be even more indeterminate than η and ξ , especially when the number of common-factors is small, so their use in the quest for parsimony may be quite illusory.

Problem VI. Rotation of axes. If η by itself is indeterminate, any rotation into an $\bar{\eta}$ remains just as indeterminate, no matter what $\bar{A}$ may result from A . Seeking meaningful factor scores merely by a correlational analysis, as implied by rotations, may be fruitless; for this further observations or experiments beyond ξ and R may be essential. Since many fundamental and useful properties of a common-factor space can be shown to exist and can be studied without specifying a particular set of reference axes [8, 9, 11], the rotation problem may have diverted attention from what can really be learned by a correlational analysis alone.

We shall actually solve Problem I for the case where the covariance matrix U^2 is not necessarily diagonal, or where the so-called β -law of deviation holds [9, p. 308]. The restriction of U^2 to a diagonal matrix turns out not to affect the existence and determinacy of possible solutions. Of our six problems, only Problem II is devoted exclusively to the special case of a diagonal U^2 .

The mathematics of factor analysis has usually been cast in terms of matrix algebra. More generally appropriate is the linear algebra of an abstract Euclidean vector space which we shall adopt in this paper.

II. THE PROBLEM OF DETERMINACY

3. *Notation for Score Matrices.* Let x_j denote the j th variable selected from the universe of content, and x_{ji} the observed score of individual i on x_j . It has been customary to regard the x_{ji} as elements of a finite real matrix of order $n \times N$. However, since this implies that the observed variables are discrete for the given population, we shall prefer instead to use a notation that will allow for both the discrete and the continuous cases, for finite and for infinite N (countable or not countable). To this end, we shall define the *score matrix* ξ and its transpose ξ' to be respectively

$$\xi = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}, \quad \xi' = [x'_1 \ x'_2 \dots x'_n]. \quad (3)$$

That is, ξ is a single column with *statistical variables* as elements, and ξ' is the same elements written as a row. We define the *inner-product* of two elements x_α and x_j to be their covariance, and denote this by $x_\alpha x'_j$. While neither x_α nor x_j is a real number, their product $x_\alpha x'_j$ is. In particular, the variance of x_j is $x_j x'_j$.

With this convention as to inner-products, we can form the product $\xi \xi'$ by the usual rule of matrix multiplication. Clearly, this is a real covariance matrix of order n ; we shall denote it by R :

$$R = \xi \xi' = [x_\alpha x'_\alpha]. \quad (4)$$

It is customary to assume all observed variances to be equal to 1:

$$x_j x'_j = 1 \quad (j = 1, 2, \dots, n). \quad (5a)$$

In this case, R in (4) is the *correlation* matrix for the variables in ξ . Convention (5a) is not essential to this paper; but we desire to restrict ourselves to the case where no diagonal element of R vanishes or all observed variances are positive:

$$x_j x'_j > 0 \quad (j = 1, 2, \dots, n). \quad (5b)$$

Since (5a) can always be attained when (5b) holds (by dividing through by standard deviations), the reader is at liberty to interpret R in this paper either as a correlation matrix or merely as a covariance matrix, according as he wishes to adopt (5a) or only (5b).

One consequence—though not the most important—of our inner-product convention is that the origins of measurement of the x_j need not be specified; the means are arbitrary. There is, however, no loss of generality in assuming the means to be zero for the problem of the present paper. When N is finite, a score matrix such as ξ in (3) can be regarded as a partitioned real matrix, each element representing a row of N real numbers. Our rule for post-multiplication of score matrices is the same as the usual one for partitioned real matrices, except that the usual result is divided through by N . Since we also wish to deal with variances and covariances for $N = \infty$, the usefulness of our modified post-multiplication rule will be apparent; it allows for general *integrals* (Stieltjes-Lebesgue) of products as well as for ordinary product-sums. This is the basic motivation for our inner-product convention. Here we shall always consider n finite. Although we shall also be interested in certain limits as $n \rightarrow \infty$, any given value of n is finite.

4. The Vector Space Concept. When dealing with score matrices, in view of our definition of an inner-product, not all rules for real matrices (i.e., matrices with real numbers as elements) need apply.

For example, we do not define the product $\xi' \xi$; but we shall also have no need for such a definition. On the other hand, if M is any real matrix of order $m \times n$, then the product $M\xi$ is defined by the ordinary law for a matrix product to yield a new column score matrix of m elements; linear combinations of variables yield variables. Similarly, if α and β are two column score matrices of n elements each, then the sum $\alpha + \beta$ is formed by the ordinary law for addition of matrices to yield a new column score matrix of n elements, for the sum of two variables is a variable. In short, column score matrices are to be pre-multiplied only by real matrices and post-multiplied only by row score matrices (in the latter case, the product being a real matrix of covariances). The remaining rules of matrix manipulations will hold for the needs of this paper.

In effect, we are regarding our statistical variables as points or elements of a *function space* (cf. 13, chap. V]. Such a space is a special case of a *Euclidean vector space* [1, p. 189]. The *norm* (length) of each of our elements is taken as its standard deviation, and the *inner-product* of any two of our elements is their covariance.

The algebra we need turns out to be completely appropriate to any finite dimensional (n finite) Euclidean vector space. It is actually more convenient to develop the algebra as for the general abstract case, and then *interpret* it for our statistical problem. Lemmas and theorems will usually be stated here then as for an n -dimensional Euclidean vector space. When we wish to analyse the general case, we shall speak, for example, of ξ as a *column of n vectors*. For our statistical case, ξ is a score matrix, or column of n statistical variables. The adaptability of the notation of matrix algebra to score matrices carries over—to the same extent—to columns of vectors from any Euclidean vector space.

One departure we shall make from a convention usually adopted for Euclidean vector spaces concerns the concept of a *zero vector*. By the zero vector, denoted by 0, is meant that vector such that $x + 0 = x$ for all vectors x in the space. The convention is usually made that all vectors with *zero norms* shall be set equal to 0. We shall prefer to denote a vector with a zero norm by 0^* , and allow it to be possibly different from 0. In our statistical case, a zero norm means a zero standard

The Determinacy of Factor Score Matrices

deviation, or a variable 0^* is *constant almost everywhere* (i.e., for almost all individuals), or equals *zero almost everywhere* if the mean is zero. It may be important in some cases to distinguish between a variable which is identically zero and one that merely has a zero variance, and we shall leave room for this distinction.

A vector with a zero norm will be called here a *null vector*. Clearly the vector 0 is a special case of a null vector. What all null vectors 0^* have in common with the vector 0 is, that if x is any vector, then the inner-product of x with 0^* vanishes: $0^*x' = 0$. This follows from Schwarz's inequality [1, p. 190]. A column of null vectors will be denoted below by θ .

A distinction between a Euclidean space and a Euclidean vector space is that in the former there is but one null vector, namely 0 . As is customary, we shall use the symbol 0 to denote several different things: the real number zero, the zero vector, a zero matrix of real numbers, and a column of zero vectors. The context should always make clear which of these is implied.

5. *The Basic Hypotheses of Factor Analysis.* Given an observed score matrix ξ , part of the basic hypothesis of the Spearman-Thurstone factor theory is that hypothetical score matrices η and ζ exist, as well as a real weight matrix A , such that: each y_k of η has a correlation of zero with each z_j of ζ , or

$$\eta\zeta' = 0, \quad (6)$$

and (1) is satisfied.

If A is of order $n \times q$, then η is composed of q variables y_k which are called *common-factors*. ζ necessarily has n elements z_j in (1), although some of these may have zero variances; these n variables we shall call here *deviant-factors*. And, following a terminology suggested in [9, p. 308], we shall say that any η and ζ satisfying (6) obey the *β -law of deviation*.

The covariance matrices of the q common-factors and the n deviant-factors will be denoted by L and U^2 respectively:

$$L = \eta\eta', \quad U^2 = \zeta\zeta'. \quad (7)$$

The Spearman-Thurstone theory requires further that ζ follow the *γ -law of deviation*, in the terminology of [9]. This means that $z_g z_g' = 0$ whenever $g \neq j$, or U^2 in (7) is a diagonal matrix. In such a case, if we denote the j th main diagonal element of U^2 by u_j^2 , we can use the conventional notation for a diagonal matrix to write:

$$U^2 = [u_1^2, u_2^2, \dots, u_n^2]. \quad (8)$$

When both the β - and γ -laws are satisfied, η and ζ are said to obey the *δ -law of deviation* [9, p. 308]. Following Thurstone, we shall then call z_j and u_j^2 the *unique-factor* and *uniqueness* respectively of x_j .

The solution to our sufficiency Problem I turns out to be essentially the same for the general β -law as for the more specialized δ -law. So we shall not restrict U^2 necessarily to be a diagonal matrix.

If L is a diagonal matrix, then the y_k are said to be orthogonal to each other; otherwise they are called oblique. We shall treat the general case of any L , diagonal or not.

It has been customary to assume the variances of the y_k all to be equal, and hence to set them all equal to 1. In some recent developments [9, 11], it appears desirable to express the y_k with unequal variances. So we shall treat the general case of arbitrary variances, and not necessarily assume the main diagonal elements of L all to be equal, nor in particular equal to 1.

The case of (1) usually studied is where A is of rank q and the elements of η are linearly independent, or L is non-singular. Since no extra work turns out to be involved thereby, we shall place no restrictions in general on the ranks of A and L . Of course, if the rank of A is r , then from the order of A it must be that $q \geq r$ and $n \geq r$.

Post-multiplying both members of (1) through in turn by η' and ζ' and using (6) and (7) show that

$$\xi\eta' = AL, \quad \xi\zeta' = U^2 \quad (9)$$

Formulae (9) are well known for the δ -law; they hold more generally for the β -law.

The basic necessary condition (2) follows from post-multiplying both members of (1) by ξ' and using (4) and (9). Condition (2) involves no score matrix of (1) directly. It is well known for the case of a diagonal U^2 , but holds again for the more general case of the β -law.

6. *The Existence of Solutions.* Hypotheses (1) and (6) by themselves do not serve to pin down q , r , or U^2 , and certainly not A and L . Even when U^2 is restricted to being a diagonal matrix, this does not improve the situation. While their procedures may assume

many forms, in effect most factor analysts begin by assuming the δ -law, and seek a diagonal U^2 that will leave $R - U^2$ Gramian.

Suppose some decision on U^2 has been reached, and that for this U^2 , the rank of $R - U^2$ is r . Factor analysts then proceed to 'factor' $R - U^2$ by any of several methods, to obtain a matrix A of order $n \times r$ and of rank r (that is, $q = r$), and a non-singular Gramian matrix L of order r , such that (9) is satisfied. The most general formulae for computing such an A and L from $R - U^2$ are given in [6, p. 12] and again in [7, p. 210]. These formulae hold whether U^2 is diagonal or not.

Our problem is: to what extent is satisfying (2) sufficient for establishing a solution to (1) that is also consonant with (6) and (7)? A general answer follows immediately from the following lemma.

Lemma 1. *Let ξ be a column of n vectors of rank p , each vector having a positive norm. Let $R = \xi\xi'$. If B and G are any real matrices such that*

$$R = \xi\xi' = BGB' \quad (10)$$

where G is a Gramian of order g and B is of order $n \times g$, then there exists a column ϕ of g vectors such that

$$\xi = B\phi, \quad \phi\phi' = G. \quad (11)$$

For the proof, we first notice that the rank of G in (10) cannot be less than that of R (the rank of a product cannot exceed that of any factor in the product), and hence must be positive. Since G is Gramian, for any finite integer f not less than the rank of G there exists a real matrix F of order $g \times f$ such that

$$FF' = G. \quad (12)$$

From (10) and (12), $R = (BF)(BF)'$, so that the rank of R is that of BF , or BF must be of rank p . Since f is the number of columns in BF , clearly $f \geq p$. Then ξ can be regarded as imbedded in a Euclidian vector space of f dimensions. Let ω_f be the column of f vectors that form an orthonormal basis for this f -space; such a basis always exists (cf. 1, p. 193). Then there is some real matrix B_f such that $\xi = B_f\omega_f$. But since $\omega_f\omega_f' = I_f$ (the identity matrix of order f), the last equality implies $R = B_fB_f'$. BF and B_f both being 'factor' matrices of R and of the same order, there exists a real orthogonal matrix W (of order f) such that $B_f = BFW$ [4, p. 71]. Therefore, if we let $\phi = FW\omega_f$, then $B\phi = B_f\omega_f = \xi$, or the first part of (11) is satisfied. Also $\phi\phi' = (FW\omega_f)(FW\omega_f)' = FF' = G$, or the second part of (11) is satisfied. Hence, a ϕ always exists for (11) and Lemma 1 is established.

As an immediate consequence of Lemma 1 we have:

Theorem 1. *If A , L , and U^2 are given such that $R = \xi\xi' = ALA' + U^2$, where L and U^2 are Gramian and no diagonal element of R vanishes, then there exist two columns of vectors, η and ζ , such that $\xi = A\eta + \zeta$ and $\eta\eta' = L'$, $\zeta\zeta' = U^2$, and $\eta\zeta' = 0$: that is, a solution to Problem I always exists.*

The proof consists of defining B and G in Lemma 1 by:

$$B = \|A I_n\|, \quad G = \left\| \begin{matrix} L & O \\ O & U^2 \end{matrix} \right\|, \quad (13)$$

where I_n is the identity matrix of order n . By direct multiplication, it is seen that $BGB' = ALA' + U^2$. Therefore, from Lemma 1 there exists a ϕ to satisfy (11), and (13). If L is of order q and U^2 of order n , then $g = q + n$. Let η be the first q elements of ϕ , and ζ the last n elements. Then

$$\phi = \left\| \begin{matrix} \eta \\ \zeta \end{matrix} \right\|, \quad \phi\phi' = \left\| \begin{matrix} \eta\eta' & \eta\zeta' \\ \zeta\eta' & \zeta\zeta' \end{matrix} \right\|. \quad (14)$$

But $\phi\phi' = G$ from (11), so comparing the second part of (13) with that of (14) shows that η and ζ satisfy (6) and (7). Furthermore, by direct multiplication according to the first parts of (13) and (14), $B\phi = A\eta + \zeta$ or—using the first part of (11)—we have verified that (1) is satisfied. This establishes the existence of η and ζ for Theorem 1.

Theorem 1 generalizes Theorem C of [7] to infinite N and to the general Euclidean vector space, and to the case where ζ is not regarded as fixed in advance.

It is of interest to inquire further *how many* solutions there are and how they are interrelated. From the proof of Lemma 1, it is obvious that ϕ cannot be uniquely determined if the rank of G exceeds the rank of R . Correspondingly, in Theorem 1, if the number of linearly independent elements in a pair of η and ζ exceeds p (the rank of R), more than one pair of solutions exist.

The Determinacy of Factor Score Matrices

Results of considerable importance are obtained by looking closely into the cases where p is less than the sum of the ranks of L and U^2 . This we shall do by *constructing* all possible solutions to Problem I. An incidental consequence will be an alternative proof of the statements in the preceding paragraph. More important, detailed information will be made available as to the relationship among alternative solutions and the implication these have for the foundations of common-factor theory. The Spearman-Thurstone theory is an important example of where p is less than the sum of the ranks of L and U^2 .

7. *A General Construction Lemma.* Lemma 1 asserts the *existence* of a solution ϕ for (11), given (10). We should now like to show how all such solutions can be *constructed*. This is done in Lemma 2.

Lemma 2. A necessary and sufficient condition that ϕ satisfy (11) in Lemma 1 is that it be of the form

$$\phi = K\xi + C\omega + \theta, \quad (15)$$

where K is a real matrix of order $g \times n$ and satisfies (16):

$$KR = GB': \quad (16)$$

C is of the order $g \times f$, as defined by (17):

$$C = (I_g - KB)F, \quad (17)$$

F being also of order $g \times f$ and satisfying (12); ω is a column of f orthonormal vectors satisfying (18):

$$C\omega\xi' = 0, \quad \omega\omega' = I_f; \quad (18)$$

and θ is a column of g null vectors satisfying:

$$B\theta = 0, \quad \theta\theta' = 0. \quad (19)$$

The proof revolves around the properties of K and the products $BK\xi$ and BC .

That a solution K to (16) always exists when R is non-singular is obvious, for then K is uniquely determined as

$$K = GB'R^{-1} \quad (|R| > 0) \quad (20)$$

When R is singular, we can lean on the theory of least-squares to see that there is not only one solution K to (16), but infinitely many. For Lemma 1 assures us that at least one ϕ exists that satisfies (11). If we expand the left number of (21) and notice from (11) that $\xi\phi' = B\phi\phi' = BG$, then we see that (21) is equivalent to (16):

$$(\phi - K\xi)\xi' = 0. \quad (21)$$

For function space, (21) is precisely the set of normal equations for the least-squares predictions of the variables in ϕ from the variables in ξ , and it is well known that a best set of regression weights K always exists, although not uniquely determined when $\xi\xi'$ is singular [13, pp. 151 f.]. The proof in [13] is given as for a function space, but clearly holds for any Euclidean vector space.

Equality (21) reveals the 'statistical' meaning of the proposed construction of ϕ in (15). $K\xi$ is the 'least-squares' estimate of ϕ from ξ , and we have to show that $C\omega + \theta$ is always the matrix of errors of estimate.

That an ω always exists to satisfy (18) should be evident. Indeed, for the first part of (18) it is sufficient to have $\omega\xi' = 0$. When our vectors are all statistical variables, ω can be defined by throwing dice or turning a roulette wheel and making the variances all equal to 1. Or ω can be taken as an orthonormalization of any set of f linearly independent scores that are uncorrelated with those of ξ .

That a θ always exists for (19) is obvious. For example, choose $\theta = 0$. This is the only choice possible if $B'B$ is non-singular, but there are many other possibilities when $B'B$ is singular, especially in a function space.

Granted that the right member of (15) always exists, we have to show its sufficiency, namely, that it satisfies (11). To this end, it will be helpful to establish Proposition 1.

Proposition 1. If E and M are real matrices such that $MEE' = 0$, then $ME = 0$. Similarly, if ψ is a column of vectors such that $M\psi\psi' = 0$ and the norm of each vector in ψ is positive, then $M\psi = 0$.

The proof of the first part of the Proposition follows from the identity $(ME)(ME)' = MEE'M'$. The right member certainly vanishes when MEE' does, and in particular its main diagonal does; but the sum of the elements of this main diagonal is the sum of squares of all the elements in ME , or each element in ME must vanish. To prove the second part, we can let $\psi = E_0\omega_0$, where ω_0 is an orthogonal basis for the space of ψ . Then $M\psi\psi' = M_0EE_0' = 0$, whence $ME_0 = 0$ from the first part of the Proposition. Therefore $ME_0\omega_0 = M\psi = 0$, or the second part is established.

Notice the requirement in the second part of Proposition 1 that the norm of each element of ψ be positive. This is to ensure that $M\psi$ should actually equal zero and not just be null.

To proceed with the proof of sufficiency for Lemma 2, let ϕ_1 and ϕ_2 be defined respectively as

$$\phi_1 = K\xi, \quad \phi_2 = C\omega, \quad (22)$$

so that (15) can be rewritten as

$$\phi = \phi_1 + \phi_2 + \theta. \quad (23)$$

Then ϕ_1 is the 'least-squares' estimate of ϕ from ξ , and ϕ_2 is our candidate for the error of estimate.

Since K always exists, so does ϕ_1 . But unlike K , ϕ_1 is uniquely determined by (16) even when R is singular. For suppose K and K^* are any two solutions to (16). Since $(K - K^*)R = 0$, Proposition 1 assures us that $(K - K^*)\xi = 0$, or $K^*\xi = K\xi = \phi_1$ even although $K \neq K^*$. Unlike ϕ_1 , ϕ_2 is quite arbitrary, being dependent on a rather arbitrary ω ; indeed ϕ_2 is fixed if and only if $C = 0$. Regardless, we shall now show that always

$$B\phi_1 = \xi, \quad B\phi_2 = 0. \quad (24)$$

Pre-multiply (16) through by B and recall (10) to see that

$$BKR = R, \quad (25)$$

or $(BK - I_n)R = 0$. Therefore, from Proposition 1 $(BK - I_n)\xi = 0$ or—expanding, and recalling the first part of (22)—the first part of (24) holds. Similarly, if we can show that $B(I_g - KB)G = 0$, then from Proposition 1 and from (17) and (12) it will follow that $BC = 0$. Expanding, we find $B(I_g - KB)G = BG - BKBG = RK' - BK RK' = RK' - RK' = 0$, the third member following from (16) and the fourth member from (25). Therefore $BC = 0$. Certainly then $BC\omega = 0$ for any ω , or the second part of (24) always holds.

The proof of (24) could be a bit shorter for non-singular R , for then $BK = I_n$ from (25). It was the case of singular R that required us to invoke Proposition 1.

We can now pre-multiply (23) through by B and use (24) and (19) to obtain $B\phi = \xi$. This verifies that any ϕ constructed according to (15) must satisfy the first part of (11). Curiously enough, according to (24), ϕ_1 also always satisfies the first part of (11). We must now see about the second part of (11).

From (22) and the first part of (18) it follows that

$$\phi_1\phi_2' = 0. \quad (26)$$

Post-multiplying each member of (23) by its own transpose and using (26) and (19) yield

$$\phi\phi' = \phi_1\phi_1' + \phi_2\phi_2'. \quad (27)$$

Post-multiplying the first part of (22) by its own transpose yields

$$\phi_1\phi_1' = KRK'. \quad (28)$$

Post-multiplying the second part of (22) by its own transpose and remembering (17) and (12) yield

$$\phi_2\phi_2' = CC' = (I_g - KB)G(I_g - B'K'). \quad (29)$$

Post-multiplying (16) through by K' shows that

$$K RK' = GB'K' = KBG, \quad (30)$$

the last member being obtained by taking the transpose of the middle member as is justified by the symmetry of the first member. Expanding the last member of (29) and using (30) and (10) yield

$$CC' = G - K RK'. \quad (31)$$

Therefore, from (27), (28), (29) and (31) we see that the second part of (11) is satisfied, or we have completed proof of the sufficiency of (15) for (11).

By capitalizing on the above algebra, the proof of necessity is now relatively rapid. Let ϕ be any solution to (11). Let ϕ_1 be as defined in (22). As we have seen, ϕ_1 is always uniquely determined by ξ , B , and G and thus is always fixed for our problem. It remains to be shown that $\phi - \phi_1$ is necessarily of the form $C\omega + \theta$.

Post-multiplying (21) through by K' and using the definition of ϕ_1 in (22), and then in turn expanding the resulting left member, show that

$$(\phi - \phi_1)\phi_1' = 0, \quad \phi\phi_1' = \phi_1\phi_1' = \phi_1\phi'. \quad (32)$$

Hence, expansion of the left member of (33) and use of (32) yield

$$(\phi - \phi_1)(\phi - \phi_1)' = \phi\phi' - \phi_1\phi_1' = G - K RK', \quad (33)$$

the last member following by recalling (11) and (28). Then from (33) and (31),

$$(\phi - \phi_1)(\phi - \phi_1)' = CC'. \quad (34)$$

The Determinacy of Factor Score Matrices

If any main diagonal element of CC' vanishes, so must the entire corresponding row of C . Let c be the number of non-vanishing rows of C , let C_c be the non-vanishing c rows of C , and let $(\phi - \phi')_c$ be the c vectors of $(\phi - \phi_1)$ with positive norms. Then from Lemma 1, there is an orthonormal ω_c such that $(\phi - \phi_1)_c = C_c \omega_c$ [in Lemma 1 replace ξ by $(\phi - \phi_1)_c$, B by C_c , G by I_c , and ϕ by ω_c]. Let θ be a column of vectors which is the same as $\phi - \phi_1$ for the $g - c$ null elements, but with zero elements elsewhere. Thus, θ satisfies the second part of (19). Let ω be any column of f orthonormal vectors containing ω_c and arranged so we can write

$$(\phi - \phi_1) = C\omega + \theta. \quad (35)$$

This can be done, for $C\omega$ contains $C_c\omega_c$ and only zeros elsewhere by virtue of the vanishing rows of C . Hence $\phi - \phi_1$ must be precisely of the form implied for $\phi_2 + \theta$ in (23).

One detail not explicitly pointed out yet is that $C\omega$ here necessarily satisfies the first part of (18); but this is precisely what (21) implies. A final detail is that θ must satisfy the first part of (19). This follows by pre-multiplying (35) through by B and using (11), (24), and (22).

The proof is thus complete as to the necessity as well sufficiency of (15) for (11), or Lemma 2 has been established.

8. *The Construction of all Solutions to Problem 1.* Definitions (13) for B and G and notation (14) for ϕ transform Lemma 2 into:

Theorem 2. A necessary and sufficient condition that η and ξ satisfy Theorem 1 is that they be of the form:

$$\eta = K_q \xi + P\omega + \theta, \quad \xi = K_n \xi - AP\omega - A\theta \quad (36)$$

where K_q and K_n are of orders $q \times n$ and $n \times n$ respectively and satisfy:

$$K_q R = LA', \quad K_n R = U^3; \quad (37)$$

P is of order $q \times f$ and satisfies:

$$PP' = L - K_q RK_q'; \quad (38)$$

ω is a column of f orthonormal vectors satisfying:

$$P\omega\xi' = 0, \quad \omega\omega' = I_f, \quad (39)$$

and θ is a column of q null vectors ($\theta\theta' = 0$).

Conditions (37) constitute but a direct restatement of (16), letting K_q and K_n be the first q and last n rows of C respectively. In (38), P is the first q rows of C in (17); the right member of (38) is the Gramian submatrix of the first q rows and columns of the right member of (31). The last n rows of C must then be equal to $-AP$, from the fact that $BC = 0$ and from the definition of B in (13). We have used this dependence of the last n rows of C on the first q in writing the coefficient of ω for ξ in (36). Similarly, the θ in (36) is the first q elements of the θ of (15), while $-A\theta$ in (36) is the last n elements of the θ of (15), according to the first part of (19); thus in Theorem 2, the analogue of the first part of (19) always holds and need not be stated explicitly. Orthonormal ω is the same in (39) as in (18), and clearly the first part of (39) is equivalent to that of (18) from the dependence now of the last n rows of C , namely AP , on the first q , or on P .

Thus Theorem 2 is but a restatement of Lemma 2 for the special case where B and G are of the form (13), and its proof has been completed.

If $P \neq 0$ in (36), clearly infinitely many solutions η can be constructed by (36), and in general also infinitely many solutions ξ , by virtue of the arbitrariness of ω . But it is also clear that the extent to which these solutions can differ among themselves, as measured by the norms of their differences, should be limited by how close the main diagonal elements of PP' are to zero. The precise nature of this limitation will be established next.

9. *The Maximal Difference among Alternative Solutions.* Given that ϕ is one solution to (11), let us seek another solution, ϕ^* , to (11) that is maximally different from ϕ in the sense that the norm of each vector in the difference $\phi - \phi^*$ is a maximum. From (23), we can write $\phi^* = \phi_1 + \phi_2^* + \theta^*$, where ϕ_1 is the same as for ϕ , but $\phi_2^* + \theta^*$ is specific to ϕ^* . Then $\phi - \phi^*$ is the same as $\phi_2 - \phi_2^*$ except for the null column $\theta - \theta^*$, so our problem is to maximize the norms of the elements of $\phi_2 - \phi_2^*$. These norms are the main diagonal elements of

$$(\phi_2 - \phi_2^*)(\phi_2 - \phi_2^*)' = \phi_2 \phi_2' - \phi_2 \phi_2^{*'} - \phi_2^{*'} \phi_2' + \phi_2^{*'} \phi_2^{*'} \quad (40)$$

Since $\phi_2 \phi_2' = \phi_2 = \phi_2^{*'} CC'$, only $\phi_2 \phi_2^{*}'$ and $\phi_2^{*'} \phi_2'$ —which have identical corresponding

diagonal elements—can vary with ϕ^* in the right of (40). Our problem reduces to maximizing the main diagonal elements of $-\phi_2\phi_2^{*'}.$ From Schwarz's inequality [1, p. 190], since ϕ_2 and ϕ_2^* have the same norms for corresponding elements, each main diagonal element of $-\phi_2\phi_2^{*'}.$ cannot be larger than the corresponding diagonal element of $\phi_2\phi_2.$ Clearly, this maximum is attainable for all diagonal elements simultaneously by letting $\phi_2^* = -\phi_2.$

To verify that $-\phi_2$ can actually constitute a column of errors of estimate, notice that if $\phi_2 = C\omega$ in (15), then we can write $\phi_2^* = C(-\omega),$ and $-\omega$ satisfies (18) just as well as ω does. According to Lemma 2, our maximally different ϕ^* is as legitimate a solution to (11) as the given ϕ is.

These results can be summarized as:

Lemma 3. If ϕ is a solution to (11) in Lemma 1, then there exists a ϕ^ that is maximally different from $\phi,$ the maximal squares of the norms of the difference column $\phi - \phi^*$ being the main diagonal element of*

$$(\phi - \phi^*)(\phi - \phi^*)' = 4CC', \quad (41)$$

where CC' is as given in (29) and (31).

The right number of (41) comes from (29) and the fact that $\phi_2 - \phi_2^* = 2\phi_2.$ The equality in (41) holds since $\phi - \phi^*$ equals $\phi_2 - \phi_2^*$ up to a null column.

Returning to Problem 1, Lemma 3 leads to:

Theorem 3. If η and ξ are a pair of solutions for Theorem 1, then there exists a maximally different pair of solutions, η^ and $\xi^*,$ such that:*

$$(\eta - \eta^*)(\eta - \eta^*)' = 4PP' = 4(L - K_r RK_r') \quad (42)$$

$$(\xi - \xi^*)(\xi - \xi^*)' = 4APP'A = 4(U^2 - K_n RK_n'), \quad (43)$$

where the real matrices of (42) and (43) are as defined in Theorem 2.

The equalities in (42) and (43) are merely expansions of (41) for case (13), using identities established for Theorem 2.

For our application to statistical variables, PP' is the covariance matrix of the errors of estimate of η from $\xi,$ and $APP'A$ is the corresponding covariance matrix of the errors of ξ from $\xi.$ According to the main diagonals of (42) and (43), the maximal variance of the difference between two alternative solutions for a given factor—common or deviant—is 4 times the variance of the errors of estimate of that factor from $\xi.$ This would lead one to suspect that ξ will have to afford very close estimates of a given pair η and ξ if all other possible pairs of solutions are not to differ materially from this given one. Such a suspicion is well-founded, as the numerical examples of the next section show.

10. Numerical Examples of the Difference Possible between Alternative Solutions for the same Factor. From now on, we shall use the special terminology of our statistical problem, although the formulae all remain appropriate to any Euclidean vector space.

Instead of dealing with variances of differences and variances of errors of estimate, it will be convenient to convert them here into correlation coefficients. (In the terminology of Euclidean vector space, we shall treat now the *cosines* of the angles between vectors.) As is well known, if two variables have equal variances $\sigma^2,$ and if ρ^* is the correlation between the two variables, then the variance of their difference equals $2\sigma^2(1 - \rho^*).$ Furthermore, if one of these variables has multiple correlation ρ on a set of predictors, then the variance of the errors of estimate equals $\sigma^2\rho^2.$ Using these two facts in conjunction with the main diagonals of (42) and (43) shows that for each factor, common or deviant, $2\sigma^2(1 - \rho^*) = 4\sigma^2\rho^2,$ whence

$$\rho^* = 2\rho^2 - 1, \quad (44)$$

where ρ^* is the correlation between an element y of η (z of ξ) with the corresponding element of the maximally different η^* ($\xi^*).$ and ρ is the multiple correlation of y (z) on $\xi.$

While ρ^2 in (44) can vary between 0 and 1, ρ^* can vary more widely between -1 and $1.$ It may be useful to exhibit specific numerical values for the relation (44). These are displayed in Table 1.

The Determinacy of Factor Score Matrices

Table 1. The Minimal Correlation (ρ^*) always attainable between Two Alternative Solutions for the Same Factor (Common or Deviant), as a Function of the Multiple Correlation (ρ) of that Factor on the Observed Scores.

ρ	.00	.30	.60	.80	.90	.95	.97	.99
ρ^*	-1.00	-.82	-.28	.28	.62	.81	.88	.96

To ensure any positive correlation at all between y and y^* (z and z^*), the multiple correlation of y (z) on ξ must exceed .71. But Table 1 reveals that even when ρ reaches .80, the corresponding ρ^* is only .28. The situation improves, but not too much, when a ρ of .90 is reached. Only when a ρ as large as .99 is attained do the alternative solutions necessarily begin to correlate by as much as .96 with each other.

In a previously published numerical example where $q = 3$ and $n = 7$, it turned out that the ρ were .90, .89, and .88 respectively for the elements of η , while they ranged from .59 to .92 for the elements of ζ [5, pp. 182 f.]. It should now be evident that having determined U^2 , A , and L for such data does not go too far towards fixing any corresponding score matrices η and ζ .

Godfrey Thomson reported that the ρ for nine 'primary traits' in Thurstone's original analysis range from .630 to .908, and remarked that: 'Those correlations do not look so bad' [15, p. 339]. If we look at these ρ again from the point of view of ρ^* , it seems that the sought-for traits are not very distinguishable from radically different possible alternative traits for the identical factor loadings.

11. *The Case of Non-Singular R.* When R is non-singular, some formulae above can be conveniently expressed in terms of R^{-1} . From (37), we have K_q and K_n uniquely defined by

$$K_q = LA'R^{-1}, \quad K_n = U^2R^{-1}. \quad (45)$$

If we let η_1 and ζ_1 be the predicted values of η and ζ respectively from ξ , and η_2 and ζ_2 the corresponding errors of prediction up to a null column, then from (36) and (45) we can write:

$$\eta_1 = LA'R^{-1}\xi, \quad \zeta_1 = U^2R^{-1}\xi, \quad (46)$$

$$\eta_2 = P\omega, \quad \zeta_2 = -AP\omega. \quad (47)$$

From (46) and (4), we find

$$\eta_1\eta_1' = LA'R^{-1}AL, \quad \zeta_1\zeta_1' = U^2R^{-1}U^2. \quad (48)$$

Now, from (27),

$$\eta\eta' = \eta_1\eta_1' + \eta_2\eta_2'. \quad (49)$$

Then, from (49) and the first parts of (7), (47), and (48), we have the identity:

$$\eta_2\eta_2' = PP' = L - LA'R^{-1}AL. \quad (50)$$

Similarly, since

$$\zeta\zeta' = \zeta_1\zeta_1' + \zeta_2\zeta_2', \quad (51)$$

we have the identity

$$\zeta_2\zeta_2' = APP'A' = U^2 - U^2R^{-1}U^2. \quad (52)$$

An interesting result of (52) is that, if U^2 is non-singular, we derive an identity for R^{-1} :

$$R^{-1} = U^{-2} - U^{-2}APP'A'U^{-2}. \quad (53)$$

This is obtained by pre- and post-multiplying (52) through by U^{-2} .

12. *The Case where L and U^2 are Non-Singular: The Matrix Q.* If both L and U^2 are non-singular, further interesting and useful special identities emerge. First, let us notice the important fact that U^2 non-singular implies R non-singular, which may be worth stating as a theorem:

Theorem 3. If R , U^2 , and $R - U^2$ are each Gramian, and if U^2 is non-singular, then R is non-singular.

This follows from the theorem that the rank of sum of two Gramian matrices cannot be smaller than the rank of either term in the sum [4, p. 73], and clearly $R = (R - U^2) + U^2$. In our case, $R - U^2 = ALA'$, according to (2).

Notice that L non-singular by itself says nothing at all about the rank of R , but the rank of ALA' does (being a lower bound to the rank of R).

Since we now assume both L and U^2 to possess inverses, we are at liberty to define a real symmetric matrix Q of order q by:

$$Q = L^{-1} + A'U^{-2}A. \quad (54)$$

With this definition, we immediately establish the following theorem:

Theorem 4. The matrix Q as defined in (54) is Gramian and non-singular. Furthermore,

$$Q^{-1} = PP' = L - LA'R^{-1}AL, \quad (55)$$

where P is as defined in (38), or Q^{-1} is the covariance matrix of the estimates of the common-factor scores from ξ .

Perhaps as easy a way as any to prove Theorem 4 is to multiply the right member of (54) by the last member of (55) and see that the product (in either order) reduces to Iq , using the fact that $ALA' = R - U^2$.

That the inverse of R in the right of (55) exists is assured by Theorem 3. Indeed, we can re-write (53) as:

$$R^{-1} = U^{-2} - U^{-2}AQ^{-1}A'U^{-2}. \quad (56)$$

Identity (56) was established before in [3, pp. 91 f.] for the special case of the δ -law (U^2 diagonal) and where $L = I_q$ and $q = r$. For that case, it often serves as an economical way of inverting R by inverting a usually much smaller matrix Q [5]. The identity holds more generally for the β -law, for $L \neq I_q$, and for $q \neq r$, although it need not lead to parsimonious calculations in the general case.

Pre-multiplying (56) through by LA' and recalling (45) and (54) establishes a further useful identity:

$$K_q = LA'R^{-1} = Q^{-1}A'U^{-2}. \quad (57)$$

For the δ -law and assuming $L = I_q$ and $q = r$, (57) was established previously in [3] and [14], being generalized to $L \neq I_r$ in [12, p. 279]. We have now further generalized it to the β -law and $q \neq r$.

Another way of arriving at (57) is to note from (26) and (14) that

$$\eta_1\zeta'_1 + \eta_2\zeta'_2 = 0. \quad (58)$$

Hence, from (58), (45), (46), and (47),

$$K_q U^2 = LA'R^{-1}U^2 = PP'A'. \quad (59)$$

For (59), only the assumption that R^{-1} exists is needed. If in addition both U^2 and L are non-singular, we can immediately derive (57) from (59), using (55).

The error covariance matrices can now be written as

$$\eta_2\eta'_2 = Q^{-1}, \quad \zeta_2\zeta'_2 = AQ^{-1}A', \quad (60)$$

according to (50), (52), and (55). No diagonal element of Q^{-1} can vanish from the Gramian nature of Q^{-1} . Similarly, at least r diagonal elements of $AQ^{-1}A'$ must be positive, for the rank of $AQ^{-1}A'$ is the rank of A , again from the Gramian nature of Q^{-1} [4, p. 75]. Since these diagonal elements are the variances of the errors of prediction from ξ , we can state:

Theorem 5. For finite n , if both L and U^2 are non-singular, all common-factors and at least r deviant-factors have positive variances of errors of prediction from ξ .

For the sake of completeness, we may remark that identities for the covariance matrices of the predictions are, respectively:

$$\eta_1\eta'_1 = L - Q^{-1}, \quad \zeta_1\zeta'_1 = U^2 - AQ^{-1}A', \quad (61)$$

as follows from (49), (51), (7), and (60).

13. *The Infinite Universe of Content and the Infinite Population.* A multiple correlation coefficient cannot decrease as the number of predictors increases: in general the coefficient increases. Hence, as n increases, it may be expected that the multiple correlation ρ on ξ will increase in general for each element of η and ζ . Although for any finite n , and when no main diagonal element vanishes in PP' or $APP'A'$, none of the correlations ρ can equal unity, this does not prevent their limits from being unity as $n \rightarrow \infty$. In particular, though L and U^2 may be non-singular for all n , so that at least $q+r$ factor scores are not perfectly predictable for any n according to Theorem 5, nevertheless all factor scores can possibly have their respective $\rho \rightarrow 1$ as $n \rightarrow \infty$.

The Determinacy of Factor Score Matrices

In developing a psychological theory, any set of n observed variables to be factored will ordinarily be regarded as but a sample from an infinitely large universe of content. Since we now wish to study what may happen as $n \rightarrow \infty$, we can no longer in general consider N to be possibly finite. Of special interest is where R is non-singular for all n . In order for R to be non-singular in (4), clearly it is necessary that N be not less than n . (Indeed, if the linear restriction is laid on the x_{ji} that their means over i be zero, we must have $N-1 \geq n$.)

In any event, finite N is not the case of greatest interest. Not only does this restrict n to being finite for non-singular R , but it prevents the x_j from being *continuous* random variables (in particular, from having a theoretical normal distribution). In a quest for scientifically meaningful factors, finite N occurs usually for but a *sample* from an infinite population. Sampling problems cannot be studied profitably until the population problems are defined. Sampling considerations are secondary to the basic problems with which we are concerned here. Our problems remain even if there is no sampling error due to people, and do not arise because of such error.

From now on we hypothesize that $N = \infty$; our basic population may be countably or not countably infinite. It was largely for this purpose that the conventions about score matrices were introduced in § 3 above. We now wish to consider what may happen as n increases.

In the general case, q and r are functions of n , so we shall write q_n and r_n ($n = 1, 2, \dots$). If U^2 , A , and L are determined for a given n , the elements of these matrices may possibly have to be modified—apart from the enlargement of their orders with n , q_n , and r_n —to keep $R = U^2$ Gramian as n increases. (In the special case where q_n and r_n can remain constant for all n sufficiently large, say for all $n > n'$, then the L determined for $n = n'$ can possibly remain constant thereafter, as can also the first n' rows and columns of U^2 and the elements in the first n' rows of A .)

To emphasize their dependence on n , we shall now attach the subscript n also to the matrices defined previously in this paper. In the case of the column score matrix ξ_n of the first n observed variables, the dependence on n takes the form only of enlarging the number of variables in the matrix and not in changing those already in the matrix for a smaller value of n . Similarly, the observed intercorrelation matrix R_n depends on n only in that the number of rows and columns increases with n ; but once an element occurs for a given n , it stays fixed in value as n increases. Not so for R_n^{-1} ; here each element changes value in general as n increases [cf. 8, p. 281]. In consequence, each element of K_{nn} and K_{nn} in (45) may change with n , as well as of P_n in (50) and (52). Correspondingly, η_n and ζ_n may have each element vary as n increases, according to the construction in (36). In particular, each of the elements of each of the column factor components may vary: $\eta_{1n}, \eta_{2n}, \zeta_{1n}, \zeta_{2n}$.

14. *Conditions for Determinacy.* What we have to consider in place of (2) is a *sequence* of equalities:

$$R_n = A_n L_n A_n' + U_n^2 \quad (n = 1, 2, \dots) \quad (62)$$

for which we seek columns η_n and ζ_n such that, in place of (1),

$$\xi_n = A_n \eta_n + \zeta_n \quad (n = 1, 2, \dots) \quad (63)$$

and (6) and (7) are satisfied for each n (or better, for all n sufficiently large).

We shall say that a factor score (element of a column score matrix) associated with the sequence (62) is *determinate* if and only if $\lim_{n \rightarrow \infty} \rho_n = 1$ for this factor. Clearly, a necessary and sufficient condition for the k th common-factor to be determinate is that the k th diagonal element of $\eta_{2n} \eta_{2n}' = P_n P_n'$ tend to zero as $n \rightarrow \infty$. Similarly, a necessary and sufficient condition for the j th deviant-factor to be determinate is that the j th diagonal element of $\zeta_{2n} \zeta_{2n}' = A_n P_n P_n' A_n'$ tend to zero as $n \rightarrow \infty$. But the vanishing of the diagonal element of a Gramian matrix necessitates the vanishing of the entire corresponding row and column. The vanishing of all diagonal elements implies the vanishing of the entire Gramian matrix. Hence, the theorem:

Theorem 6. A necessary and sufficient condition that all common-factors be determinate for (62) is that the left member of (64) exist and (64) hold:

$$\lim_{n \rightarrow \infty} P_n P_n' = 0. \quad (64)$$

The corresponding condition for deviant-factors is that the left member of (65) exist and (65) hold:

$$\lim_{n \rightarrow \infty} A_n P_n P'_n A'_n = 0. \quad (65)$$

If R_n is non-singular for all n sufficiently large, then (64) is equivalent to

$$\lim_{n \rightarrow \infty} (L_n - L_n A'_n R_n^{-1} A_n L_n) = 0, \quad (66)$$

and (65) is equivalent to

$$\lim_{n \rightarrow \infty} (U_n^2 - U_n^2 R_n^{-1} U_n^2) = 0. \quad (67)$$

If L_n and U_n^2 are non-singular for all n sufficiently large, then (64) and (65) are equivalent to the respective parts of (68):

$$\lim_{n \rightarrow \infty} Q_n^{-1} = 0, \quad \lim_{n \rightarrow \infty} A_n Q_n^{-1} A'_n = 0, \quad (68)$$

where Q_n is as defined in (54) for all n sufficiently large.

The equivalence of the above conditions to each other follows from (50), (52), and (60).

Evidently, from (44), if $\rho_n \rightarrow 1$ then $\rho_n^* \rightarrow 1$ for the same factor. Or, from (42) and (43), if (68) holds then all variances of differences between alternative solutions tend to zero as $n \rightarrow \infty$. Hence Theorem 7:

Theorem 7. A sequence (62) can lead to at most one column of determinate common-factors, say η_∞ , and at most one column of determinate deviant-factors, say ζ_∞ . That is, all alternative solutions for a determinate factor are equal up to null columns.

15. *The Case of Diagonal U_n^2 (the δ -law).* Although we have used the terms 'common' and 'deviant' for the η_n and ζ_n , there is nothing basic in the algebra thus far that distinguishes between the two types of factors. The β -law as expressed by (6) is essentially symmetric in η and ζ . Indeed, our algebra would be symmetric throughout if we had considered not η but $A\eta$ in (1) or (63). Essential asymmetry begins to occur when further restrictions beyond the β -law are laid on either η or ζ . The δ -law, or combining the γ -law with the β -law, leads to a particularly fundamental type of asymmetry, for then diagonal U_n^2 becomes intimately related to the main diagonal of R_n^{-1} .

Let u_{jn}^2 be the j th diagonal element of U_n^2 . If u_{jn}^2 is always positive, then clearly in order for (67) to hold the non-diagonal elements in the j th row and column of R_n^{-1} must tend to vanish as n increases. This implies:

Theorem 8. If all deviant-factors are determinate and if U_n^2 is diagonal and non-singular for all n sufficiently large, then R_n^{-1} must tend to a diagonal matrix as $n \rightarrow \infty$.

Notice that Theorem 8 assumes U_n to be non-singular for all n sufficiently large, so R_n^{-1} exists for all n sufficiently large according to Theorem 3. That the main diagonal of R_n^{-1} must always have a limit as $n \rightarrow \infty$ is a consequence of multiple-correlation theory, as will be marked in more detail in the next section below.

The tendency of R_n^{-1} to a diagonal matrix has been discussed from a point of view closely related to the present one in [3, p. 95]. As has been pointed out in [8, p. 282 and pp. 294 f.), this tendency can serve as a basis for a test of the hypothesis that there is a useful δ -law structure at all to the universe of content. If R_n^{-1} does not tend to a diagonal matrix, then there can be no determinate unique-factor scores, nor hence a determinate common-factor space.

Our condition for the non-diagonal elements of R_n^{-1} to tend to vanish is somewhat different from—though very closely related to—that in [3]. We require in Theorem 8 only that the unique-factors be determinate and have non-vanishing uniqueness for all n sufficiently large.

III. SUPPLEMENTARY PROBLEMS

16. *A Solution to Problem II (Communalities).* By considering only what happens to the main diagonal of R_n^{-1} as n increases, we arrive at a solution to the problem of communalities of the δ -law. As is well known, the j th main diagonal element of R_n^{-1} is the reciprocal

The Determinacy of Factor Score Matrices

of the variance of the errors of estimating x_j from the remaining $n-1$ variables from the observed universe. Let σ_{jn}^2 denote this variance [the 'partial anti-norm' of x_j in the terminology of (8)].

Inspecting the main diagonals of (67) shows that for each j :

$$\lim_{n \rightarrow \infty} u_{jn}^2 \left(1 - \frac{u_{jn}^2}{\sigma_{jn}^2} \right) = 0. \quad (69)$$

Since a variance of errors of estimate cannot increase as the number of predictors increases, σ_{jn}^2 is a monotonely decreasing function of n for all j , so there always exists $\lim_{n \rightarrow \infty} \sigma_{jn}^2$ (the 'total anti-norm' of x_j), which we shall denote by $\sigma_{j\infty}^2$. (We have used this existence proposition in Theorem 8; it implies that each element of the main diagonal of R_n^{-1} always tends to a limit, finite or infinite, as $n \rightarrow \infty$.)

Considering (69) and its source in (52), we can state:

Theorem 9. *If there exists $\lim_{n \rightarrow \infty} u_{jn}^2$, say $u_{j\infty}^2$, then a necessary and sufficient condition that the j th unique-factor associated with sequence (62) be determinate for diagonal $U_{j\infty}^2$ is that either $u_{j\infty}^2 = 0$ or $u_{j\infty}^2 = \sigma_{j\infty}^2$.*

The restriction of R_n to being non-singular has been omitted in Theorem 9, for it has been shown in [8, p. 293] that $u_{jn}^2 \leq \sigma_{jn}^2$ even when R_n is singular. [That $u_{jn}^2 \leq \sigma_{jn}^2$ when R_n is non-singular follows also from (52) for diagonal U_n^2 .] Hence, if $\sigma_{jn}^2 = 0$ for any n , $u_{jn}^2 = 0$ for that n , and Theorem 9 holds also for this singular case.

According to Theorem 9, if the main diagonal of R_n is to be modified by subtraction of a non-singular U_n^2 to leave only the contribution of common-factors in (62), the choice is limited only to a U_n^2 which will tend to whatever the main diagonal of R_n^{-1} tends. This provides a solution to the problem of communalities. *From the point of view of determinacy of factors, the ideal choice of uniquenesses is $\sigma_{j\infty}^2$ ($j = 1, 2, \dots$).*

A striking feature of this solution for uniquenesses (or equivalently, for communalities) is that it lays no restrictions on the common-factor space. Nothing is said about the rank of $R_n - U_n^2$ for any n , nor about the determinacy or location of the common-factors. The limit of the rank of the common-factor space can be finite or infinite. This may help explain why past attempts to solve the problem of communalities for finite n have not been successful. Apparently, only lower bounds to communalities are universally possible for finite n [cf. 10]. The ρ_{jn}^2 afford such bounds, and must tend to the actual communalities if the unique-factors are to be determinate with positive uniquenesses.

17. Implications for Problem III: the Scientific Meaning of Factor Analysis. Since Spearman's original analysis of the problem when $q_n = r_n = 1$ and U_n^2 non-singular ($n = 1, 2, \dots$), it has been generally recognized that factor scores cannot be estimated exactly from the observed content scores.

For example, Holzinger and Harman state: "Since the total number of factors (both common and unique) exceeds the number of observed variables, the value of any particular factor for a given individual cannot be obtained by direct solution but can only be estimated from the observed values of the variables. The best prediction, in the least-square sense, is that obtained by the ordinary regression method" [12, p. 264]. Thurstone seems to regard this estimation problem as largely a practical one; and his position is that "the problem of appraising an individual as to each of several factors is best solved in the practical situation by using a test composite for each factor" [16, p. 515].

These authors are focusing on a common-factor score matrix η_n . They apparently take it for granted not only that η_n exists, but also that it is uniquely determined by U_n^2 , A_n , and L_n , even though it can be only approximated as a linear function of the variables in ξ_n . They do not discuss what can be said about η_n beyond its predictability from ξ_n .

Thomson has gone more intensively into this problem, and Ledermann has generalized his conclusions. Assuming one pair of solutions η_n and ζ_n exists, they show how further solutions can be obtained from this pair for the same A_n and U_n^2 , assuming $L_n = I_r$ [cf. 15, pp. 371 f.]. While recognizing that "usually the accuracy is very low indeed" for predicting factors from the observed ξ_n [15, p. 336], Thomson has not considered the relationship of ρ^* to ρ as was done in §§ 9-10 above. As remarked above, Thomson considered that ρ between .630 and .908 "do not look so bad", whereas from (44) this implies ρ^* between -.206 and .650.

It seems safe to say that there has been considerable complacency about being able to ascribe particular scientific meaning to η_n and ζ_n on the basis of A_n , L_n , and U_n^2 , even if the

factors are not determined closely for the given n . It has been especially true with respect to common-factors that they have been *named* according to the content of the observed variables that have 'high' loadings on them in A_n .

Thurstone has further insisted that "*it is a fundamental criterion of a valid method of isolating primary abilities that the weights of the primary abilities for a test must remain invariant when it is moved from one test battery to another test battery*" [16, p. 361, italics his]. Here again is expressed the belief that A_n delimits some particular η_n with no reference to the ρ_n or ρ_n^* .

It appears from the relation of ρ_n^* to ρ_n (for each intended factor) that the predictability of factors from ξ_n is not merely a practical problem. If ρ_n^* is low, it raises the question of what it is that is being estimated in the first place; instead of only *one* 'primary trait' there are many widely different variables associated with a given profile of loadings. If the scientific usefulness of factor analysis is to rest on its finding factor scores that can serve as meaningful reference axes, surely these reference axes should be defined by the analysis as being distinct from alternative axes.

No matter how well an A_n satisfies a given criterion of rotation, and no matter how constant some of its rows may be from test battery to test battery, a particular set of meaningful reference axes is still not defined if ρ_n does not tend to unity for these axes.

If in practice ρ_n does not get to be well above .90, say, for most of the factors, so that ρ_n^* is generally above .60, it is hard to see that any meaningful reference axes have been established by computing A_n , L_n , and U_n^2 , regardless of what arithmetical criteria and manipulations are used.

Attention may be focused on the problem whether *in principle* a given universe of data admits of an approximately determinate set of common- and deviant-factors. Empirical evidence is scanty in published studies, since the ρ_n are not usually computed. The few instances of recorded ρ_n noted above are not very encouraging for the hypothesis of determinacy. The new radex approach to factor analysis [9] does not require determinacy in order for practical and theoretical use to be made of its analysis of an R_n . Not so the Spearman-Thurstone type of theory, which for mental tests is in effect a neo-faculty theory of psychology. Unless the factors or 'faculties' are pinned down, they can be used neither in theory nor in practice; hence, if determinacy is unobtainable from the observed universe of ξ , a correlational analysis of R_n by itself is not adequate for the Spearman-Thurstone type of theory. A more direct and experimental approach to the hypothesized factors may be required to test and verify their nature. If more direct observations on the η_n and ξ_n cannot be made than statistical analysis of R_n and ξ_n , the Spearman-Thurstone approach may have to be discarded for lack of determinacy of its factor scores.

18. *Problem IV: 'Inverted' Factor Analysis.* The proposal has often been made to 'factor' people rather than variables, especially when N is finite and smaller than n [cf. 15, chaps. XIII and XIV]. Discussing such a proposal only for finite N and n clearly ignores the problem of determinacy of factor scores. Artifacts due to finiteness of the sample data certainly occur in the algebra of the usual discussions. The variables can no longer theoretically have normal distributions, as already remarked, nor any other type of continuous distribution.

There is nothing mathematically wrong with the 'inverted' problem of factoring people, provided the proper framework is specified. For determinacy, in general n must be infinite. One could consider what happens as $N \rightarrow \infty$ analogously to our present treatment. This requires specifying, somehow, inner-products for the Euclidean vector space implied, the vectors corresponding to people now rather than to variables. How to specify these seems to be as yet an unsolved empirical problem.

The most general approach may ultimately be not to 'factor' either people or variables, but to start out by assuming both N and n infinite, and to regard the totality of theoretically possible observations as defining a *Hilbert space*, rather than a finite-dimensional vector space. Our present theory started with finite n , and considered what happened as $n \rightarrow \infty$. This implicitly assumes some kind of frequency distribution of variables *over people*. A related concept of a frequency function for variables occurs explicitly in the simplex subtheory of the radex approach to factor analysis [11]. A more general formulation is under way by the writer in terms of Hilbert space and measure theory that may simultaneously take care of direct and inverted factor theory in a manner that seems natural for the problem of factor analysis, and avoiding the artifacts and indeterminacies of finite data.

Neither N nor n should ordinarily be regarded as finite at the outset in a fundamental theory for mental tests, say, or for other substantive areas for factor analysis. Starting with finite n in our present paper may be regarded as an incompleteness of our present theory. Starting with a finite N for an 'inverted' factor theory is an analogous sign of incompleteness, apart from the other problems involved.

The Determinacy of Factor Score Matrices

19. *Problem V: Second-Order Common-Factors.* The solution to Problem I has important implications for the concept of 'second-order' factors. It has been suggested that when $q = r$ and $L \neq I_r$, or oblique common-factors are used, it may sometimes be meaningful to 'factor' these r common-factors in turn [16, chap. XVIII]. The procedure would be to seek a diagonal matrix, say D , such that $L - D^2$ would be Gramian and of rank less than r ; say the rank would be s where $s < r$. The assumption is that thereby s 'second-order' common-factors and $r - s$ 'second-order' unique-factors are determined which underly the original r common-factor.

Now, Problem I applies to the second-order factors as well as to the first-order ones. If $|D^2| > 0$, then the solution to Problem I shows that infinitely many sets of second-order factors will satisfy any given rotation in this 'second-order' space. Only as $r \rightarrow \infty$ can the embarrassment of too many solutions possibly disappear.

However, advocates of second-order factors usually also advocate having r finite and relatively small. Requiring r to become indefinitely large would contradict a basic hypothesis of this school of thought. If r is to be finite, this requires either: (a) giving up the notion of second-order factors, or (b) hypothesizing that the second-order factors scores are *in principle indeterminate* for any given rotation of axes in the second-order space. In case (b), the question might be raised as to the psychological meaning or usefulness of the concept of second-order factors as reference axes.

20. *Problem VI: Rotation of Axes.* One of the currently outstanding questions of common-factor theory is that of rotation of axes. This concerns the term $A\eta$ in (1). Given that A and η satisfy (1), let T be any real non-singular matrix of order q and let $\bar{A}$, $\bar{\eta}$, and $\bar{L}$ be defined by

$$\bar{A} = AT^{-1}, \quad \bar{\eta} = T\eta, \quad \bar{L} = TLT'. \quad (70)$$

Then clearly

$$\bar{A}\bar{\eta} = A\eta, \quad (71)$$

or $\bar{A}$ and $\bar{\eta}$ satisfy (1). Furthermore, $\bar{\eta}$ satisfies (6), and also (7) and (9) with $\bar{A}$ and $\bar{L}$ in place of A and L .

$\bar{\eta}$ is called a 'rotation' of η , being an 'orthogonal rotation' if T is an orthogonal matrix. $\bar{\eta}$ is a solution to Problem I when A and L are replaced by $\bar{A}$ and $\bar{L}$. Hence there are as many $\bar{\eta}$ corresponding to a given rotation matrix T as there are η for (1) as is. The same indeterminacy problem of scores holds for all rotations if only an indirect analysis is made via A and T . Since the criteria for rotations used by different schools of thought are in terms of the parameters or elements of $\bar{A}$, these again ignore the determinacy problem of scores. Although we arrived at a solution to Problem II, or the determination of U^2 , this does not help at all to solve Problem VI or the fixing of T when A is given.

Enlarging the perspective of Problem VI to include also Problem I suggests that perhaps undue emphasis has been placed on Problem VI in the past. If reference scores $\bar{\eta}$ are not going to be pinned down thereby anyway, according to the solution to Problem I, why bother about rules for $\bar{A}$ and T ? Such rules alone may not enable one to make much use of the implied factors, or to distinguish them from quite different factors.

As remarked in § 17, perhaps factor analysis has strained too much to try to isolate meaningful factors by means of a correlational analysis alone. Many important and useful things can be learned about the structure of ξ and R from a correlational analysis alone, without recourse to reference factor scores, as radex theory has shown [9, 11]. Perhaps we should be content with this latter type of structural information until we can make direct observations and experiments beyond ξ and R in order to determine $\bar{\eta}$.

REFERENCES

1. Birkhoff, Garrett, and MacLane, Saunders (1953). *A Survey of Modern Algebra* (rev. ed.). New York: Macmillan.
2. Dwyer, P. S. (1940). 'The evaluation of multiple and partial correlation coefficients from the factorial matrix.' *Psychometrika*, V, 211-232.
3. Guttman, Louis (1940). 'Multiple rectilinear prediction and the resolution into components.' *Psychometrika*, V, 75-99.
4. Guttman, Louis (1942). 'Properties of Gramian matrices', Chap. II, *The Prediction of Quantitative Variates by Factor Analysis* (Ph.D. Thesis, Univ. Minnesota).
5. Guttman, Louis, and Cohen, Jozef (1943). 'Multiple rectilinear prediction and the resolution into components: II.' *Psychometrika*, VIII, 169-183.
6. Guttman, Louis (1944). 'General theory and methods for matrix factoring.' *Psychometrika*, IX, 1-16.
7. Guttman, Louis (1952). 'Multiple group methods for common-factor analysis: their basis, computation, and interpretation.' *Psychometrika*, XVII, 209-222.
8. Guttman, Louis (1953). 'Image theory for the structure of quantitative variates.' *Psychometrika*, XVIII, 277-296.
9. Guttman, Louis (1954). 'A new approach to factor analysis: the radex', in Lazarsfeld, Paul F. (ed.), *Mathematical Thinking in the Social Sciences*. Glencoe, Ill.: The Free Press.
10. Guttman, Louis (1954). 'Some necessary conditions for common-factor analysis.' *Psychometrika* XIX, 149-161.
11. Guttman, Louis (1955). 'A generalized simplex for factor analysis.' *Psychometrika*, XX: (in press).
12. Holzinger, K. J., and Harman, H. H. (1941). *Factor Analysis*. Chicago: Univ. Chicago Press.
13. Jackson, D. H. (1930). 'The theory of approximation.' *American Mathematical Society Colloquium Publications*, II. New York: Am. Math. Society.
14. Ledermann, W. (1939). 'On a shortened method of estimation of mental factors by regression.' *Psychometrika*, IV, 109-116.
15. Thomson, G. H. (1950). *The Factorial Analysis of Human Ability*. London: Univ. London Press.
16. Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago: Univ. Chicago Press.

ERRATA

The editor regrets that the following errors have been made in printing Professor Guttman's article on 'An Additive Metric from all the Principal Components of a Perfect Scale' in the preceding issue of this *Journal*, (VIII, pp. 17-24).

P. 22, equation (3). *Instead of "δ" read "δ_{ij}".*

P. 23, equation (15). *The right member should read: $-\frac{1}{\eta_a^2} f_0 x_{a0}$.*

P. 23, equation (16). *In the right member instead of " $x_{a i, i+1}$ " read " $x_{a, i+1}$ ".*

P. 23, equation (20). *Instead of " p_{ij} " read " p_{0j} ".*

P. 24, equation (22). *Instead of " p_{ij} " read " p_{1j} ".*

FACTOR ROTATION FOR PROPORTIONAL PROFILES: ANALYTICAL SOLUTION AND AN EXAMPLE

By R. B. CATTELL and A. K. S. CATTELL

Laboratory of Personality Assessment and Group Behaviour, University of Illinois

I. *Existing Attempts at Unique Factor Resolution.* II. *The Rationale of Proportional Profiles.*
III. *History of Empirical Explorations of the Problem.* IV. *Analytical Solution for Pro-*
portional Profiles with Orthogonal Axes. V. *A Worked Example.* VI. *Summary.*

I. EXISTING ATTEMPTS AT UNIQUE FACTOR RESOLUTION

Factor analysts may be divided—rather sadly divided—into (a) those for whom factors are mainly mathematical utilities, specific in reference to a particular matrix or mathematical process and quite restricted in psychological meaning (if any), and (b) those for whom factor analysis is a means of discovering and dealing with scientific concepts that persist through all matrices concerned with similar phenomena. With increasing realization of the scope of factor analysis in the social sciences, and with the accumulation of invariant results [8], the emphasis on factors as scientific hypothesis has fortunately begun to prevail. The need to expand this conceptual usage of factors makes it all the more urgent to discover a technique for determining in any factorial investigation that unique rotational solution which will correspond to the presumed inherent structure in nature.

Some seven distinct principles in use for gaining a unique, meaningful solution were listed in an earlier article [5]; six of them have been enlarged upon in a book [7] which also reviewed critically the principle of criterion rotation [12], but not the proposal of Sandler [21] which, though ingenious, merely passes the responsibility from *R*- to *Q*-technique and asks the blind to lead the blind. Of existing methods none seems to the writer to have been so successful as 'simple structure'; for, in the first place, it has a reasonably clear theoretical basis—the principle of parsimony; and, secondly, it has proved itself pragmatically, that is, so far as pragmatism can ever be acceptable. The pragmatic proof consists in having yielded psychologically meaningful factors more often than other methods have done, and in providing an impressive accumulation of comparisons [8, 13, 15] that have revealed invariance of factor pattern from one study to another—a criterion less tainted by subjectivity than the first.

It would, however, be a mistake to assume that simple structure can be accepted as a final satisfactory solution to the rotation problem. It suffers from at least three shortcomings: (1) different users do not define and use it according to a single acceptable definition; (2) the labour and skill required to obtain simple structure is such that half the people who set out to obtain it fall far short of doing so; and (3) the procedure has certain theoretical weaknesses. The second difficulty promises to be removed by electronic computers, thanks to the methods devised by Carroll [3], Neuhaus and Wrigley (Quartimax) [19], Saunders [22], Thurstone [26] and Tucker [28], though recent experience shows that they are not altogether fool-proof and may still require to be supplemented by craftsmanship. However, it is the theoretical inadequacies of the procedure that are important and will concern us here.

The main irremediable theoretical failure arises from the fact that, if a factor is to be located by simple structure, it must leave an appreciable fraction of the variables unloaded, i.e., in a discernible hyperplane. Now it is conceivable, and indeed it often happens, that a factor will affect *all* the variables in a matrix. This can occur first because (a) the variables have not been chosen with sufficient catholicity and with due regard for ensuring 'hyper-plane

stuff' [7], e.g., by using the 'personality sphere' concept [6] or some similar principle of variable sampling: this is an avoidable fault of design, though judging by its frequency it is hard to escape. Secondly, it may occur because (b) there exist factors which are indeed so pervasive and ubiquitous (like g in the cognitive field, or surgency-desurgency in the field of personality) that no clear hyperplane can ever be found for them. The converse but equally disconcerting finding is that in one and the same factorization, with n factors extracted, one can sometimes find more than n hyperplanes, so that there exist alternative resolutions. The senior author has found and reported such equally attractive alternatives [8, 10], though in a prolonged experience they have been rare. Other writers, e.g., Bargmann [1], maintain they are quite frequent; and, judging from the occurrence of alternative solutions by competitive psychologists, the discrimination of the better from the poorer or false hyperplane is not always easy. The principle now to be propounded considers real structure to be inherent in the data, but does not depend for resolution on the elusive hyperplane.

II. THE RATIONALE OF PROPORTIONAL PROFILES

The principle called 'parallel proportional profiles' was proposed in 1944 [3] as a method for avoiding the restrictions of simple structure. Only a brief indication of its nature therefore need be given here.

Since the assumptions and objectives of proportional profiles are intrinsically difficult to convey, as evidenced by the number of misunderstandings, certain points must be emphasized. The basic assumption is that, *if a factor corresponds to some real organic unity, then from one study to another it will retain its pattern, simultaneously raising or lowering all its loadings according to the magnitude of the role of that factor under the different experimental conditions of the second study.* No inorganic factor, a mere mathematical abstraction, would behave in this way, within the realm of orthogonal factors. The principle suggests that every factor analytic investigation should be carried out on at least two samples, under conditions differing in the *extent* to which the same psychological factors (working as independent, orthogonal influences) might be expected to be involved. We could then anticipate finding the 'true' factors by locating the unique rotational position (simultaneously in both studies) in which each factor in the first study is found to have loadings which are proportional to (or are some simple function of) those in the second: that is to say, a position should be discoverable in which the factor in the second study will have a pattern which is the same as in the first, but is stepped up or down. Obviously, if no change in the relative importance of the same factors occurs from the first to the second experiment, no rotational position can be determined; but it is assumed that, either by natural influences or by deliberate experimental control, the real factor influences will differ beyond chance from one measurement situation to the other.

In the original statement of this theorem [4], a general practical solution was not attempted, and proof was limited to demonstrating (a) that proportionality of loadings cannot be attained by any orthogonal rotation whatever of two factor matrices (with the same number of factors and variables) taken at random, i.e., that it depends on the existence of real 'organic' structure in the data, and (b) that, if *one* position exists in which the loadings of factors in one matrix are proportional to those of corresponding factors in the other, then no other pair of orthogonal rotations exist for which this is true, i.e., there is uniqueness. It is proposed to show here how the solution may be generalized to any number of dimensions and adapted to computation. To anticipate the formal mathematical development, the problem is that of finding D , a diagonal matrix containing the several distinct proportionalities by which factors alter from experiment A to experiment B, so that we may write $V_{nA} = V_{nB}D$, where V_{nA} is the rotated position of the first factor matrix, obtained from the unrotated matrix V_{oA} by an as yet unknown transformation λ_A , and λ_A is a rectangular matrix. Thus $V_{oA}\lambda_A = V_{nA}$, and similarly $V_{oB}\lambda_B = V_{nB}$.

Before developing this, we should note that the method proposed uses the principle of parsimony involved in Thurstone's simple structure in a more comprehensive fashion. In fact, it requires the principle to hold with respect to the general field of scientific observation and inference (as, philosophically, it should) instead of merely within the universe of a single experiment or matrix. It applies it to *any* two matrices (requiring that they yield the same factors), and therefore to all. Simple structure hopes, and often finds, that what is simplest in one matrix is simplest for many; but proportional profiles *require* that all the phenomena analysed in the various matrices or experiments shall be explained by the smallest number of factors, i.e., by demonstrably similar patterns. Incidentally, the definition of

'diverse conditions' could be extended, and probably with advantage, to comparisons of *R*- and *P*-technique factorizations.

The labour of arriving at the unique proportional profiles position by simultaneous trial and error in two matrices would obviously be prohibitive; but fortunately, by the nature of the method, an analytical solution is possible, at least for errorless data. In the 1944 article [4] this was indicated for two factor problems. Before proceeding to the general solution, a digression will be made on some experimental attempts to set up the problem, since (a) they illustrate the theoretical relations more clearly, (b) they may save others from the pitfalls we encountered, and (c) they will indicate ways in which the theoretical solution still requires adaptation to real conditions.

III. HISTORY OF SOME EMPIRICAL EXPLORATIONS OF THE PROBLEM

The writers felt an urgent need of the method in psychological research; and the ten-year delay in further publication has been due not merely to the pressure of other work but to certain misfortunes with empirical try-outs of the method.

It has been argued [7, 9] that factor analysis in general would benefit from more *a posteriori* reasoning, as in Burt's bottle problem [2] or Thurstone's box problem [25], i.e., taking experimental situations in which the 'factors' or 'influences' are well known and seeing how various factor analytic techniques behave. In the present problem the advantages would be didactic (a) in demonstrating to those psychologists who are more pragmatic or philosophical than mathematical that factors are more than mathematical abstractions, (b) in showing whether the accidentally or experimentally produced changes in factor variance from sample to sample are large enough to produce the effects required, and (c) in finding whether the blurring of simple proportionality by experimental and sampling errors (in addition to the systematic distortion by selection efforts) are so large as to affect the practicability of an error-free analytical solution. The last is the most important.

Three kinds of examples, additional to the completely errorless 'pure principle' case here demonstrated, have been or are in process of trial and publication: (1) a box problem similar to Thurstone's, but with orthogonal factors and with the additional hurdle of powered relations departing from the 'pure principle' used here; (2) a similar box problem with experimental errors introduced; and (3) real and natural data, which contain both experimental and sampling error, but in which the factors are controlled and known with tolerable certainty.

The two first will be described elsewhere by Haverland [16] who has shown that the principle of proportional profiles works excellently in these cases. The third is represented by two examples which will be briefly described here and one of which is fully described elsewhere [16]. In both a search was made for a situation in which (for simplicity) *two* common factors, and two only, operated on a dozen or more variables. For the first it was decided to work on growing tomato plants, since botanists believed that two factors, representing light and nutrition levels, would be found. For help in setting up this experiment, we are greatly indebted to Dr. D. Gottlieb of the University of Illinois Horticultural Department.

Two populations, containing 200 plants in each, were grown from seed; and measurements were finally made on such variables as height of plant, thickness of stem, greenness of leaf, number of leaves, etc. In one sample the variance was made twice as great in light intensities, and in the other it was increased by 3/2 in concentration of hydroponic nutrition. The factor scores, at least as ranks, were thus known for each plant. The whole design collapsed, however, when it was found after six months that nine-tenths of the common factor variance was due to one factor only—indeed there seemed a doubt as to whether any second factor should be extracted. In this extreme factor composition no orthogonal or even oblique position of rotation could be found that would give proportionality of profiles and indeed no rotation of any kind would give a satisfactory 'light intensity' factor. Anyone repeating is advised to include more variables known botanically to be appreciably responsive to light alone.

Still seeking for living and organic (rather than merely physical) examples that would nevertheless be controllable we were encouraged by the recent discovery of drives as factors [10] to suggest that hunger and thirst be factored from the usual learning and deprivation variables on two populations of rats. This work, by Haverland [16], had larger variation of food deprivation in one sample and larger variations of water deprivation in the other. The many interesting findings cannot be described here; but it is questionable whether proportional profiles gave a meaningful solution, whereas simple structure gave easily recognizable drive patterns (best oblique, but acceptable even when orthogonal). The steps likely to produce advances beyond this situation have been indicated elsewhere [15]; but even in the purely mathematical treatment given here it would seem that a further step must eventually be taken to allow for the fact that the proportionality is only approximate. If a variable is loaded by

a factor 0.5 and the variance of that factor (still orthogonal) is doubled relative to other factors the loading will not equal 1.0. If we apply the formulae of Thomson and Ledermann [24] and Thurstone [25] for predicting the correlations to be expected after the variance of the selection variable has been changed (counting the factor as the variable on which selection is made), it will be found that not all the loadings of a factor will change in the same proportion when the factor variance is changed. However, *over small ranges* the simple proportionality is near enough, and Haverland's box example with errors [16] shows that it holds well enough to give a correct solution. Accordingly we shall here present the solution brought to that level, and leave for further invention the adjustment to non-proportionality and the other practical difficulties (e.g., obliqueness) raised by the rat problem.

IV. THE ANALYTICAL SOLUTION FOR PROPORTIONAL PROFILES, WITH ORTHOGONAL AXES

We now suppose that two experiments, A and B, with the same n variables have yielded two correlation matrices and two unrotated factor matrices. The only assumptions made are (a) that the matrices are of the same order, both containing n rows and k columns, and of the same rank, k , (b) that enough variables exist for satisfactory determination of the communalities [23], (c) that they occupy the same factor space, i.e., if the matrices are placed alongside to give order $n \times 2k$, the rank will remain k , and (d) that the variances of the true influence on these variables in the two situations are not unchanged: in fact, the variance ratios for the factors must be different from unity and from each other.

Parenthetically, the method of proportional profiles should be distinguished, in objectives and methods, from the method of aligning two factors from different studies which is implicit in the formulations of Barlow and Burt [2],¹ and from Tucker's related method [27] which has the rather different objective of synthesizing two factor studies to yield a least squares oblique solution to equation [3] below. The latter furnishes no restrictions such as are required to obtain a unique solution for the rotation; hence an additional criterion, such as simple structure or the present method (adapted to the oblique case), needs to be applied to bring the two studies marching in step to a unique halting point. A similar objective—getting maximum agreement between two studies—has been pursued by Wrigley and Neuhaus [30]: their method is distinguished from proportional profiles by being applied *after* the studies, except for the provision of common variables in the two studies. Proportional profiles on the other hand, requires the stage to be set beforehand, in order to provide a unique rotational solution dependent on controlled experimentation. It does so in order that a particular relation may be sought, which is not merely one of mathematical convenience or economy, but is posited to arise from the existence of an inherent structure.

From a rotation of a given factor matrix we obtain in the (notation above):

$$V_{oA}\lambda_A = V_{nA}, \quad (1)$$

$$\text{and } V_{oB}\lambda_B = V_{nB}. \quad (2)$$

¹*Editorial Note.* In correspondence Professor Cattell has suggested that I might indicate my own views on the problems raised. The question discussed above would seem to be a special instance of the *third* case in the paper of ours that he quotes, viz., that in which the battery of tests is the same but the sample of persons different (p. 54). We dealt with it less fully than the other two, because it had already been discussed in the references cited. For the instance he has in mind, our procedure would be to extract the initial 'bipolar' factor matrices by *weighted* summation, not by simple (i.e., not by the centroid method). With the notation there used, we should obtain

$$G_1 = F_1H_1 = L_1V_1^{\frac{1}{2}}H_1 \text{ and } G_2 = F_2H_2 = L_2V_2^{\frac{1}{2}}H_2.$$

Identity of factors is defined to mean identity of their direction-cosines (p. 52). It therefore follows that $G_2 = G_1D$. Now let $F_2 = F_1T$. Then

$$F_1T = F_2 = L_2V_2^{\frac{1}{2}} = L_1V_1^{\frac{1}{2}}H_1DH_2'.$$

Premultiplying by $V_1^{-\frac{1}{2}}L_1'$, we find $T = H_1DH_2' = V_1^{-\frac{1}{2}}L_1'V_2^{\frac{1}{2}}L_2$. The matrix, T , therefore, must then itself be factorized by weighted summation, and we at once obtain H_1, D , and H_2 . (This seems better than factorizing $TT' \equiv KK'$, which would give 'canonical multipliers' rather than latent roots and vectors.) The use of weighted summation avoids the difficulty mentioned by Professor Cattell, namely, that with his formula K and K^{-1} , when independently computed, are not consistent. I may add that, even with data that do not involve experimental or sampling errors, T (Professor Cattell's K) is as a rule not symmetrical, but (as noted in my earlier paper) tends rather to be triangular: cf. Burt, C., *Brit. J. Educ. Psychol.*, IX, 1939, pp. 45–71 and refs. But in general it seems desirable to determine the group factors separately and prove their similarity rather than to postulate their similarity and rotate them accordingly.—C.B.

The positions V_{nA} and V_{nB} are by definition such that each factor column in one matrix has loadings that are a simple multiple of those in the corresponding column of the other, assuming corresponding factors arranged in the same order. The transformation from one to the other, involved in the essential *principle of proportional profiles*, is thus simply represented by a $k \times k$ diagonal matrix D , such that:

$$V_{nA} = V_{nB}D. \quad (3)$$

Substituting from (1) and (2) into (3), we have:

$$V_{oA}\lambda_A = V_{oB}\lambda_B D, \quad (4)$$

or, keeping solely to orthogonal transformations,

$$V_{oA} = V_{oB}\lambda_B D\lambda'_A \quad (5)$$

(since $\lambda' = \lambda^{-1}$, when λ is an orthogonal matrix).

To handle the term on the right, equation (5) may more conveniently be re-written

$$V_{oA} = V_{oB}K, \quad (6)$$

$$\text{where } K = \lambda_B D\lambda'_A. \quad (7)$$

Now V_{oA} and V_{oB} are the given factor matrices, defined as to number of factors, loadings, etc., by the experimental process. Consequently they provide the empirical basis necessary for calculating K and the values derived from it. Assuming K calculated,¹ let us see how to decompose it into the values required. We first multiply K by its transpose, obtaining $KK' = \lambda_B D\lambda'_A \lambda_A D\lambda'_B = \lambda_B D^2 \lambda'_B$.

This expression has the typical form for the latent roots and latent vectors of a symmetrical matrix [29], D^2 being a diagonal matrix containing the latent roots and λ_B containing the latent vectors. Consequently, after extracting the latent roots and vectors in the usual way, we can obtain D , by taking the square roots of the elements in D^2 . With λ_B thus determined we can obtain V_{nB} ; and with D determined we can obtain V_{nA} by (3). This suffices to determine the position; but to solve completely by obtaining λ_A also it is computationally easier² to use

$$K'K = \lambda_A D\lambda'_B \lambda_B D\lambda'_A = \lambda_A D^2 \lambda'_A.$$

The solutions for λ_A and λ_B by these procedures are unique, thus confirming our earlier two-space geometrical proof of the uniqueness of the proportional profiles position [3]. It will be noted that a unique solution is not possible if our assumption is broken that the variances due to the true influences (factors) in the two situations do *not* remain unchanged. Thus if one (or more) of the diagonal values in D are unity or have the same value, one (or more) of the factors cannot be uniquely rotated.

V. A WORKED EXAMPLE

The working procedures will now be illustrated by an artificial problem involving seven variables and three factors. It has been so constructed as (1) to be free of sampling and experimental error, (2) to be susceptible of solution by the usual simple structure method, and (3) to have known proportionalities between the corresponding factors (though we shall not assume this knowledge in achieving our solution).

¹ The computation for fallible, sample-affected data has to be based on least squares. This involves minimizing the total of the squared entries in the difference matrix $V_{oA} - V_{oB}K$. This can be verified by writing these in terms of their elements, taking partial derivatives with respect to the elements of K , setting these equal to zero, and solving the resultant equations for the elements of K . This approach was suggested by Rao [20]. Turnbull and Aitken [29, p. 173] demonstrate this approximation by general methods. It has been used for related purposes by Gibson [14], Lubin [17], Mosier [18], and Tucker [27]. Gibson has suggested (personal communication) a rider whereby K can be calculated from V_{oB} and R_A only, and factored by Hotelling's iterative procedure, i.e., "by an implicit factoring which tends to select from R_A those dimensions which are represented in V_{oB} ." However this leaves the communality problem out of consideration. Unfortunately, for fallible experimental data the solutions for K are not symmetrical, and $V_{oA} = V_{oB}K$ yields a different result from $V_{oB} = V_{oA}K^{-1}$. K^{-1} may be obtained from the expression $(V_{oA}'V_{oA})^{-1}V_{oA}'V_{oB}$. The two K values can be computed as estimates and KK' calculated. This will not be an exactly symmetrical matrix; but can be made so by averaging the corresponding elements on opposite sides of the diagonal. Tucker's device, using mutual regression [27], is perhaps somewhat neater.

² Another solution for λ_A follows from [7], which gives: $\lambda_A = K'\lambda_B D^{-1}$. Nil results may be checked by $\lambda_B = K\lambda_A D^{-1}$.

Factor Rotation for Proportional Profiles

Relatively small differences of variance, such as might exist in practical examples, were chosen, such that the factor loadings in the second matrix are respectively 2/3rds, 9/8ths, and 6/5ths of those for corresponding variables in the first. To give a local habitation and a name to the problem, we can imagine that the correlation matrix represents relations between seven tests involving the manipulation of mechanical puzzles, and that success is determined by factors of mechanical aptitude (F_1), reasoning ability (F_2), and manual dexterity (F_3). The two matrices may represent data from boys and girls respectively, F_1 having less variance in the boys and F_2 and 3 in the girls. Let us begin, as it were, from behind the scenes with the two final, rotated matrices, representing the simple structure and the functionally meaningful position just described. These are set out in Table I.

TABLE I. FACTOR MATRICES

(As known to be structured)

A. Boys					B. Girls				
Tests	F_{1nA}	F_{2nA}	F_{3nA}	h^2_A	Tests	F_{1nB}	F_{2nB}	F_{3nB}	h^2_B
1	.60	-.40	-.10	.53	1	.40	-.45	-.12	.37
2	-.50	-.80	.10	.90	2	-.33	-.90	.12	.93
3	-.10	.80	.00	.65	3	-.17	.90	.00	.84
4	.00	.50	-.60	.61	4	.00	.56	-.72	.83
5	.90	-.10	.00	.82	5	.60	-.11	.00	.37
6	-.60	.00	.70	.85	6	-.40	.00	.84	.87
7	.05	.00	.80	.64	7	.03	.00	.92	.92

For uniformity we have labelled the factors the F_{nA} 's and the F_{nB} 's. It will be observed that the required proportionalities exist between the loading in each pair of columns, and that each factor in each matrix has a position of tolerable good simple structure, having three or more variables in the hyperplane for each factor (generally within ± 0.12 but in one instance 0.17); and each variable with at least one hyperplane loading [25].

The correlation matrices reconstructed from these two factor matrices, by taking the inner products of the rows ($V_{nA}V_{nA}'$), are shown in Table II.

TABLE II. OBTAINED CORRELATION MATRICES

R_A (Boys)								R_B (Girls)							
Tests	1	2	3	4	5	6	7	Tests	1	2	3	4	5	6	7
1	—							1	—						
2	.01	—						2	.26	—					
3	-.38	-.59	—					3	-.47	-.75	—				
4	-.14	-.46	.40	—				4	-.17	-.59	.50	—			
5	.58	-.37	-.17	-.05	—			5	.29	-.10	-.20	-.06	—		
6	-.43	.37	.06	-.42	-.53	—		6	-.26	.23	.07	-.60	-.24	—	
7	-.05	.06	-.01	-.48	.04	.53	—	7	-.10	.11	-.01	-.69	.02	.79	—

Taking these correlation matrices as our starting point, let us try to arrive at a meaningful rotated solution (a) by the new principle of proportional profiles and (b) by rotation for simple structure. A centroid factorization of the correlation matrices yields the following unrotated factor matrices (Table III).

TABLE III. UNROTATED FACTOR MATRICES

V_{oA} (Boys)					V_{oB} (Girls)				
Tests	F_{oA_1}	F_{oA_2}	F_{oA_3}	h^2	Tests	F_{oB_1}	F_{oB_2}	F_{oB_3}	h^2
1	34	-66	01	55	1	31	51	18	39
2	59	29	-68	89	2	71	32	-58	94
3	-63	25	45	66	3	-69	-60	14	86
4	-75	-19	-07	60	4	-88	21	05	82
5	22	-77	45	84	5	15	29	51	37
6	34	86	11	87	6	49	-78	-19	88
7	50	37	51	65	7	63	-70	22	94

Let us first solve from this point by the proportional profiles method From (6) we know

$$V_{oA} = V_{oB}K,$$

$$\text{whence } K = V_{oB}^{-1}V_{oA}. \quad (8)$$

Our next step is to get K by equation (8). However, since V_{oB} is not a square matrix, it cannot have an inverse.¹ But a device is possible which we may call the symmetrizing method. This has the advantage that it is certain to yield D ; whereas the trimming method (see footnote 2) sometimes fails. By this approach we aim first at a symmetric matrix S , such that $S = V_{oB}'V_{oB}$. In this calculation the whole V_{oB} and V_{oB} is used, yielding, in this case, a 3×3 matrix. Note that, since $S^{-1} = V_{oB}^{-1}(V_{oB})^{-1}$, and $K = V_{oB}^{-1}V_{oA} = V_{oB}^{-1}(V_{oB})^{-1}V_{oB}V_{oA}$, therefore $K = S^{-1}V_{oB}'V_{oA}$.

This use of the square non-singular matrix S thus permits us to solve² for K ; and we then obtain:

$$S = \begin{bmatrix} .4342 & .0396 & .2225 \\ .0396 & .5169 & .0379 \\ .2225 & .0379 & 1.4743 \end{bmatrix} \quad K = \begin{bmatrix} .8638 & .0070 & .0162 \\ .0854 & -.8293 & -.3971 \\ .1035 & -1.0279 & 1.0563 \end{bmatrix} \quad KK' = \begin{bmatrix} .7465 & .0615 & .0993 \\ .0615 & .8527 & .4418 \\ .0993 & .4418 & 2.1831 \end{bmatrix}$$

The latent roots and vectors were obtained on the electronic computer by the method of Truman Kelley, adapted and reported by Wrigley and Neuhaus [30]:

$$D^2 = \begin{bmatrix} 2.33 & & \\ & .76 & \\ & & .70 \end{bmatrix} \quad D = \begin{bmatrix} 1.53 & & \\ & .87 & \\ & & .84 \end{bmatrix} \quad \lambda_B = \begin{bmatrix} .0714 & .8048 & .5893 \\ .2895 & .5487 & -.7843 \\ .9545 & -.2266 & .1938 \end{bmatrix}$$

¹ Equation (8) and those at the end of the paragraph cannot validly be written because they involve the imaginary concept of the inverse of a rectangular matrix. Since we eventually advocate a different approach, it may be asked why this initial lution, with its inadequate 'matrix trimming' device, is retained. It is retained because it is logically more direct and easier for non-mathematicians to follow. For, although a rectangular V_{oB}^{-1} cannot be used, a solution could evidently be obtained by trimming it to a square matrix, since the transformation applied to k variables (rows) is applied equally to all; at least this would be so with non-fallible data, though with experimental and sampling errors no such sample of rows could be truly representative of the whole matrix. There the only satisfactory solution is the least squares fit for the whole matrix (see footnote 1). Even with exact data ambiguities arise from this squaring of the rectangle or 'trimming', as will be seen below. But for maximum illumination it has seemed best to carry through both the cruder 'trimming' and the neater 'symmetrizing' procedures for comparison, though the former will be carried out only in footnotes, since it is not advocated in practice.

By the direct 'trimming' method we first trim V_{oB} and V_{oA} to three rows each and then take the inverse of the first (see (6)) as follows:

$$V_{oB}^{-1} = \begin{bmatrix} -13.49 & -7.98 & -15.72 \\ 13.38 & 7.46 & 13.68 \\ -9.13 & -7.38 & -11.70 \end{bmatrix}$$

which, post-multiplied by the square matrix of V_{oA} (trimmed), becomes:

$$K = \begin{bmatrix} .609 & 2.657 & -1.782 \\ .328 & -3.248 & 1.221 \\ -.090 & .960 & -.335 \end{bmatrix}$$

² As will be seen, this result differs decidedly from the value for K found from the 'trimmed' matrices.

Factor Rotation for Proportional Rotation

The values for V_{nA} and V_{nB} are given by eqns. (2) and (3); and the results are shown in Table IV. λ_A can be reached, as indicated above, either by calculating $K'K$ and proceeding therefrom as from KK' (which will also give a check on D^2) or by eqn. (1), p. 86.

TABLE IV. PROPORTIONAL PROFILES SOLUTION
(Symmetrizing Approach)

$V_{nA} = V_{nB}D$ (Boys)				$V_{nB} = V_{oB}\lambda_B$ (Girls)			
Tests	F_{nA_1}	F_{nA_2}	F_{nA_3}	Tests	F_{nB_1}	F_{nB_2}	F_{nB_3}
1	.53	-.42	-.15	1	.34	-.48	-.18
2	-.63	-.77	.05	2	-.41	-.88	.06
3	-.14	.80	.08	3	-.09	.92	.09
4	.08	.52	-.56	4	.05	.60	-.67
5	.89	-.14	-.03	5	.58	-.16	-.04
6	-.57	.01	.72	6	-.37	.01	.86
7	.08	.07	.80	7	.05	.07	.96

Except for errors from rounding, these values are the same as in Table I. Even in F^1 and F_2 , where the approximation is rougher, one has no difficulty in recognizing the similar patterns. A further example since worked by Haverland with improved mechanics of computing furnished the original values correct to four decimal places [16]. Furthermore, the values for D —1.53, 0.87, and 0.84—are close enough for all practical purposes¹ to the

TABLE V. PROPORTIONAL PROFILES SOLUTION
(‘Trimming Approach’)

V_{nA} (Boys)				V_{nB} (Girls)			
Tests	F_{nA_1}	F_{nA_2}	F_{nA_3}	Tests	F_{nB_1}	F_{nB_2}	F_{nB_3}
1	.58	-.46	.01	1	.32	-.52	-.12
2	-.58	-.74	-.01	2	-.46	-.85	.12
3	-.00	.81	.00	3	-.05	.92	.00
4	.16	.48	-.59	4	.10	.53	-.73
5	.88	-.15	.20	5	.57	-.20	.00
6	-.73	.06	.57	6	-.40	.09	.85
7	-.12	.01	.80	7	.01	.02	.97

The agreement is almost as good as by the other approach; but the D matrix failed, and the V_{nA} matrix had to be calculated through $K'K$. The D matrix gave, not the known ratios, but 0.89, 4.79 and 0.00! presumably taking only three rows of a matrix is likely to reduce its rank, since four of the seven variables must be linear combinations of the others; and we are unlikely to hit by chance on those that would preserve the rank. Accordingly, since ‘symmetrizing’ gives more immediately what is required and avoids the need to find a pseudo-inverse, it seems the better procedure.

original, namely, 1.50, 0.89, and 0.83; and the closeness of these ratios to unity has provided a severer test of the method than we hope will normally be demanded.

Now let us compare with the above proportional profiles determinations the simple structure position as reached by the usual trial and error search. This is done primarily to

¹ It is interesting to compare this result with that found by the ‘trimmed’ matrix K . The V_{nA} and V_{nB} values so found are:

compare the relative computational times required and the role of ambiguity and approximation in the two methods. A blind rotation for simple structure was carried out from Table III by an experienced worker. Rotations proceeded steadily to a satisfactory structure, obtained after six over-all rotations for V_{nA} , and five for V_{nB} (Table VI).

TABLE VI. SIMPLE STRUCTURE

V_{nA} (Boys)				V_{nB} (Girls)			
Tests	F_{nA_1}	F_{nA_2}	F_{nA_3}	Tests	F_{nB_1}	F_{nB_2}	F_{nB_3}
1	.65	-.35	-.06	1	.40	-.47	-.08
2	-.14	.80	-.04	2	-.36	-.90	.07
3	-.45	-.83	.11	3	-.16	.90	.00
4	.00	.45	-.63	4	.11	.56	-.71
5	.92	.01	.08	5	.59	-.11	.06
6	-.65	-.01	.67	6	-.50	.02	.79
7	-.01	.06	.81	7	-.09	.00	.96

Blind rotation has thus succeeded in finding the original simple structure, though others *might* find an alternative structure. In their agreement with the true position (Table I) the simple structure (Table V) and the proportional profiles methods (Tables IV and V) are equally good. But in the wider applicability and in the analytical solution the proportional profiles solution can claim superiority, at least when restriction to the orthogonal case is overcome.

VI. SUMMARY

1. The assumption is made that, when (and only when) factors correspond to organic influences inherent in the data, will their loadings alter as a whole, by ratios peculiar to each, from one situation to another in which the factor as a whole is differently involved.

2. It has been shown in the orthogonal case that, when this relation exists, it can be found only by one unique paired rotation of the two factor matrices, and that this position can be discovered with certainty by an analytic solution.

3. An example has been worked on errorless data, in which the solution was made to be simultaneously the simple structure and the proportional profiles position. Re-working from the correlation matrix, this position was correctly found, without ambiguity, by two methods. The former required about seven hours of computation and drawing; the latter about two hours of tape-punching and ten minutes of electronic computing.

4. Work in progress confirms that the analytic solution can be obtained with great accuracy for errorless data, and also, with acceptable accuracy, for a box problem with introduced error. However, an actual psychological experiment [16], in which control was attempted of the factors believed to account for the variance, gave poor agreement of proportional profiles with simple structure and the presumed psychological structure.

5. Even without this example, it would be evident that the method awaits the following improvements for its effective use. (i) Adaptation to oblique axes; additional restrictions will be necessary to obtain a unique solution in this case. (ii) Allowance for the fact that, as implied in the work of Thomson and Ledermann [24], variance change will not produce a simple proportional change in the loadings of all variables, but a change which is partly a function of the individual variable loading. (iii) Allowance in individual loadings for the fact that the experimenter will often be compelled to produce his experimental change of factor variance by selecting on a variable highly loaded in the factor, instead of on the factor itself.

REFERENCES

1. Bargmann, R. (1954). 'Signifikanzuntersuchungen der Einfachen Struktur in der Faktoren-Analyse.' Sonderdruck. Physica-Verlag, Würzburg.
2. Barlow, J. A. and Burt, C. (1954). 'The identification of factors from different experiments.' *Brit. J. Stat. Psychol.*, VII, 52-53.
3. Carroll, J. B. (1953). 'An analytical solution for attaining simple structure in factor analysis.' *Psychometrika*, XVIII, 23-28.
4. Cattell, R. B. (1944). 'Parallel proportional profiles and other principles for determining the choice of factors by rotation.' *Psychometrika*, IX, 267-283.
5. Cattell, R. B. (1946). 'Simple structure in relation to some alternative factorizations of the personality sphere.' *J. Gen. Psychol.*, XXXV, 225-238.
6. Cattell, R. B. (1946). *The Description and Measurement of Personality*. New York : World Book Co.
7. Cattell, R. B. (1952). *Factor Analysis : An Introduction and Manual for the Psychologist and Social Scientist*. New York : Harper.
8. Cattell, R. B. (1955). 'The principal replicated factors discovered in objective personality tests.' *J. Abnorm. Soc. Psychol.*, L, 291-314.
9. Cattell, R. B. (1955). 'Growing points in factor analysis.' *Austral. J. Psychol.*, VI, 105-140.
10. Cattell, R. B. and Miller, A. (1952). 'A confirmation of the ergic and self-sentiment structures among dynamic traits attitude variables by R-technique.' *Brit. J. Psychol.*, XLI, 280-294.
11. Cattell, R. B. and Gruen, W. (1955). 'Measurement of the primary personality factors in children by means of objective tests.' *J. of Person.*, XXIII, 460-478.
12. Eysenck, H. J. (1950). 'Criterion rotation, an application of the hypothetico-deductive method to factor analysis.' *Psychol. Rev.*, LVII, 38-65.
13. French, J. W. (1953). 'The description of personality measurements in terms of rotated factors.' Princeton: Educational Testing Service.
14. Gibson, W. A. (1953). 'A least squares solution for case IV of the law of comparative judgment.' *Psychometrika*, XVIII, 15-21.
15. Goodman, C. H. (1943). 'A factor analysis of Thurstone's sixteen primary mental abilities tests.' *Psychometrika*, VIII, 141-151.
16. Haverland, E. M. (1954). 'The application of an analytical solution for proportional profiles rotation to a box problem and to the drive structure in rats.' Ph.D. Thesis, University of Illinois Library, Urbana, Ill.
17. Lubin, A. (1950). 'A note on "criterion analysis".' *Psychol. Rev.*, LVII, 54-57.
18. Mosier, C. I. (1939). 'Determining a simple structure when loadings for certain tests are known.' *Psychometrika*, IV, 149-162.
19. Neuhaus, J. O., and Wrigley, C. (1954). 'The quartimax method : an analytical approach to orthogonal simple structure.' *Brit. J. Statist. Psychol.*, VII, 88-92.
20. Rao, C. R. (1952). 'Advanced statistical methods in biometric research.' New York: Wiley.
21. Sandler, J. (1952). 'A technique for facilitating the rotation of factor axes, based on equivalence between persons and tests.' *Psychometrika*, XVII, 223-229.
22. Saunders, D. R. (1953). 'An analytic method for rotation to orthogonal simple structure.' Princeton : Educational Testing Service Rept.
23. Thomson, G. H. (1952). 'The Factorial Analysis of Human Ability.' London: Univ. London Press. 5th edn.
24. Thomson G. H., and Ledermann, W. (1939). 'The influence of multivariate selection on the factorial analysis of ability.' *Brit. J. Psychol., Gen. Sect.*, XXIX, 288-306.
25. Thurstone, L. L. (1947). 'Multiple Factor Analysis.' Chicago: Univ. Chicago Press.
26. Thurstone, L. L. (1953). 'Analytical method for simple structure.' *Adv. Publ.*, No 6, Psychom. Lab., Chapel Hill, N.C.
27. Tucker, L. R. (1951). 'A method of synthesis for factor analysis studies.' Washington: Dept. of the Army, Adjutant General's office, P.R.S. Rept. No. 984.
28. Tucker, L. R. (1953). 'The objective definition of simple structure in linear factor analysis.' *Psychometrika*, XX, 209-226.
29. Turnbull, H. W., and Aitken, A. C. (1932). 'An introduction to the Theory of Canonical Matrices.' London : Blackie & Son.
30. Wrigley, C. F., and Neuhaus, J. O. 'The matching of two sets of factors.' (In press.)
31. Wrigley, C. F., and Neuhaus, J. O. 'The use of an electronic computer in principal axes factor analyses.' *J. Educ. Psychol.* (In press.)

MECHANISMS INVOLVED IN GROUP PRESSURES ON DEVIATE-MEMBERS

By HERBERT A. SIMON and HAROLD GUETZKOW
Carnegie Institute of Technology

I. *Problem.* II. *Variables and Equations of the Deviate-Member Model.* III. *Comparison of Deviate-Member and Aggregate Models.* IV. *Cohesiveness and Rejection.* V. *Summary.*

I. PROBLEM

Leon Festinger and his associates have stated and tested a number of propositions about communication processes in small groups [1]. The purpose of this paper is to carry a step further the synthesis of a segment of these propositions into an interrelated system. First, we shall translate some of Festinger's propositions into a more rigorous form, and examine the correspondence between our equations and his verbal statements. Secondly, we shall examine an empirical study by Schachter [2] in its bearing upon our model.

For many purposes it is convenient to describe a group as though it were an aggregate which may be characterized by merely averaging or summing the individual characteristics of all its members. But as we attempt to explore the internal processes within groups, we find it useful to differentiate one member of the group (e.g., the leader) from the others.¹ Many of the propositions Festinger has formulated for group pressures belong to the first type; they are concerned entirely with aspects of group behaviour that can be treated as averages or aggregates for the group as a whole. Elsewhere [3] we have examined this part of Festinger's theoretical scheme. The present paper will undertake a parallel analysis of the remainder of Festinger's system—those propositions that involve specific reference to a particular member, *A*, of the group. The member selected for attention is a *deviate-member*; and the principal variables in the system represent the attitudes and responses toward the deviate-member of the remaining members. In a later section we shall comment on the relation between the two parts of Festinger's system, and between the models we have constructed to represent them.

Festinger states five hypotheses about communication resulting from pressures toward uniformity that describe the behaviour of deviate-members. Of these two contain references to the communication among individuals in a group; one is concerned with other potential membership groups; the remainder refer to changes in composition of the group, e.g., through pushing members out of it. Our procedure will be to set forth our system, and then to relate the model to Festinger's verbal hypotheses.

II. VARIABLES AND EQUATIONS OF THE DEVIATE-MEMBER MODEL

The five deviate-member hypotheses of Festinger—those numbered 2a, 2b, 2c, 4a, and 4c—are stated in terms of eight variables in addition to time. Consider a group of *n* members and a particular member, *A*. The variables refer to averages over the group *excluding A*.

R: The *relevance* to the group of uniformity in respect of the topic under discussion.

PA(t): The average *Pressure* upon members of the group (excluding *A*) to communicate to *A* at time *t*.

¹ The importance of the distinction has been recognized by the economist in applying index numbers in his analysis of economic behaviours. We speak of an 'index of wholesale prices' only if we are interested just in the general movement of prices, and if all prices tend to move together. If either of these conditions fails, the index becomes correspondingly meaningless. Thus, it is a pragmatic question as to when we can employ an aggregate model, or when we must formulate a more complete model in which the definition of the variables involves less aggregation.

Mechanisms Involved in Group Pressures on Deviate-Members

$DA(t)$: The *Discrepancy* between the opinion of A and the average opinion of the rest of the group, as perceived by the rest at time t .

$CA(t)$: The *cohesiveness* of the group with A —that is, the average intensity of the desire to retain A as a member of the group at time t .

$LA(t)$: The *perceived* receptivity ('listening') of A to influence—that is, the average estimate by the other members of his willingness to be influenced by them at time t .

$UA(t)$: The average pressure toward uniformity of the group with A , as felt by the group at time t .

$C(t)$: The *cohesiveness* of the group, exclusive of the deviate member—that is, the average intensity of the desire of the other members to remain in the group.

We now hypothesize the following system of equations:

$$P_A(t) = P_A[DA(t), UA(t)]; \text{ (see footnote (1)).} \quad (1.1)$$

$$\frac{dLA(t)}{dt} = \phi[DA(t), LA(t)]. \quad (1.2)$$

$$\frac{dCA(t)}{dt} = g_A[DA(t), UA(t), CA(t), LA(t)]. \quad (1.3)$$

$$U_A(t) = U_A[CA(t), R, C]. \quad (1.4)$$

Since Festinger has defined cohesiveness as 'the resultant of all the forces acting on members to remain in the group', the level of $C(t)$ at which members are indifferent as to the choice between remaining in or leaving the group could be taken as a natural zero for this cohesiveness variable. Analogously, $CA(t)$ may be defined as the average *net* attitude of the group (other than A) towards A 's remaining in the group. Then positive CA would mean a net preference for his remaining, negative CA would mean *rejection*—a net preference for his leaving.

The four equations constituting our model postulate that the changes in P_A and U_A are instantaneous, while the changes in LA and CA take place gradually through time, as functions of the variables appearing on the right-hand sides of their respective equations. A discussion of the nature of such systems of algebraic and differential equations is presented in the appendix.

We are now ready to examine Festinger's five hypotheses concerned with the deviate member:

Hypothesis 2a: The pressure to communicate about 'item x ' to a particular member of the group will increase as the discrepancy in opinion between that member and the communicator increases.

This hypothesis states a relationship between DA and P_A , as defined in our equation (1.1) and asserts that $\frac{\partial P_A}{\partial DA} > 0$.

Hypothesis 2b: The force to communicate about 'item x ' to a particular person will decrease to the extent that he is perceived as not a member of the group or to the extent that he is not wanted as a member of the group.

This hypothesis states a relationship between P_A and CA , via U_A (eqns. (1.1) and (1.4)) and asserts that $\frac{\partial P_A}{\partial U_A} > 0$, $\frac{\partial U_A}{\partial CA} > 0$.

Hypothesis 2c: the force to communicate 'item x ' to a particular member will increase the more it is perceived that the communication will change that member's opinion in the desired direction.

The present system of equations does not translate the full content of this proposition, in particular that part of it which refers to the 'desired direction' of change in opinion. A closely related proposition stated by Schachter [2] plays an important role in his analysis. Schachter makes the assumption 'that perceived difference increases with discussion'

¹ Read: The magnitude of the pressure upon group members to communicate to A at time t is a function of the discrepancy of A 's opinion from the rest of the group, and of the pressure toward uniformity.

when the actual discrepancy remains constant [2, p. 201]. In the present system of equations, the mechanism that is postulated to produce this kind of behaviour is slightly different. We assume that, as discussion proceeds, the perceived receptivity of A to influence (L_A) declines if the perceived discrepancy of opinion persists. This is stated in equation (1.2) with $\frac{\partial \phi}{\partial D_A} < 0$, $\frac{\partial \phi}{\partial L_A} < 0$. Further, we assume that the lower is L_A , the lower is the equilibrium

value of C_A . This is expressed by equation (3.3), with $\frac{\partial g_A}{\partial C_A} < 0$, $\frac{\partial g_A}{\partial L_A} > 0$.

By means of these additional assumptions, a high persistent value of D_A depresses L_A (eqn. 1.2); this reduces C_A (eqn. 1.3), with a consequent reduction in U_A (eqn. 1.4), and simultaneously in P_A (eqn. 1.1). Thus, if the perceived receptivity of A to influence declines, so also will the pressure to communicate to him.

Hypothesis 4a: The tendency to change the composition of the psychological group (pushing members out of the group) increases as the perceived discrepancy in opinion increases.

This hypothesis relates C_A to D_A (eqn. 1.3). It asserts that $\frac{\partial g_A}{\partial D_A} < 0$.

Hypothesis 4b: When non-conformity exists, the tendency to change the composition of the psychological group increases as the cohesiveness of the group increases and as the relevance of the issue to the group increases.

This relates C_A to R and C via U_A and asserts that $\frac{\partial g_A}{\partial U_A} < 0$; $\frac{\partial U_A}{\partial R} > 0$; $\frac{\partial U_A}{\partial C} > 0$ (eqns. (1.3) and (1.4)).

Our system now consists of the equations (1.1) to (1.4) together with certain assumptions about the signs of partial derivatives of the functions that appear in them. For convenience, we collect the latter assumptions.

$P_D > 0$	(1.1 a)	$P_U > 0$	(1.1 b)		
$\phi_P < 0$	(1.2 a)	$\phi_L < 0$	(1.2 b)		
$g_D < 0$	(1.3 a)	$g_U < 0$	(1.3 b)	$g_C < 0$	(1.3 c)
$U_C > 0$	(1.4 a)	$U_{R'} > 0$	(1.4 b)	$U_{C'} > 0$	(1.4 c)
				$g_L > 0$	(1.3 d)

For notational simplicity, we have suppressed the subscript A , and employed subscripts to denote partial differentiation, designating the two variables without subscripts, R and C , by R' and C' respectively.

Deviation and Rejection: The Schachter Experiment. A group of persons, including several confederates or 'paid participants', is brought together for discussion. The variable D_A , where A is one of the confederates (the 'deviate'), is determined by the (pre-arranged) behaviour of the confederate, and hence remains constant. Different initial values of R and C are induced by the instructions given at the outset of the experiment. The cohesiveness of the groups was checked at the end of the experiment and was found constant (see (2) Table II, p. 194); R is assumed to remain constant for the forty-five minutes of the discussion. Direct measurements of C and C_A were also made at the end of the trial. During its course, measurements were made of $P_A(t)$. These were average values over ten minute intervals.

The Schachter experiment provides us with a test of the postulated model in the following respects:

(1) We can test whether differences in the parameter values, C and R , and D_A produced the predicted differences in the terminal values of C_A .

(2) We can test whether the time paths of $P_A(t)$ are consistent with those predicted by the model.¹

Eliminating U_A by equation (1.4) and regarding D_A , R , and C as parameters, we obtain the

equations:
$$\frac{dL_A}{dt} = \phi(D_A, L_A). \quad (2.1)$$

$$\frac{dC_A}{dt} = g_A[D_A, U_A(C_A, R, C), C_A, L_A]. \quad (2.2)$$

¹ Schachter, in his paper, attempts to derive inferences about the time paths of the variables from assumptions about the signs of certain partial derivatives. Because the basic equations of his model are algebraic rather than differential equations, he can do this only by introducing the *ad hoc* assumption 'that perceived difference increases with discussion time'. It is difficult to know precisely what assumptions underlie the time paths displayed in his Fig. 3. By mathematizing the assumption, as was done following Hypothesis 2c, it is possible to know exactly what assumptions are made.

Mechanisms Involved in Group Pressures on Deviate-Members

Consider now the direction field of the system in the (C_A, L_A) plane, with the parameters held constant.¹ Since C_A does not appear in (3.5) we have:

$$\left[\frac{\delta L_A}{\delta C_A} \right]_{\phi=0} = 0. \quad (2.3)$$

$$\left[\frac{\delta C_A}{\delta L_A} \right]_{g_A=0} = - \frac{g_L}{g_U U_C + g_C} > 0. \quad (2.4)$$

There are grounds for a plausible assumption which enables us to say that these curves have a particular shape—namely, that $\frac{\delta C_A}{\delta L_A}$ (eqn. 2.4) approaches zero for very large and very small values of

C_A . Then the direction field will have the general form depicted in Fig. 1. The argument for this is as follows: Equation (2.2) describes the change in the cohesiveness of the group with A . It seems reasonable that this mechanism is subject to saturation, i.e., that when C_A reaches very high levels, it will not be driven much higher by further decreases in D_A . Likewise, when C_A reaches very low levels, it will not be driven much lower by further increases in D_A . Typical paths for the system are shown from initial positions A, B, C , and D . There is a single stable equilibrium for the system at E .

Now, referring to equation (2.2), and remembering our assumptions as to signs, we find that an increase in any one or more of the parameters D_A, R , and C corresponds to a shift to the left of $g_A = 0$; while a decrease produces a shift to the right of $g_A = 0$. Similarly, from (2.1) an increase in D_A means a downward shift of $\phi = 0$ and a decrease in D_A an upward shift. It follows that an increase in one or more of the parameters shifts E to the left (and downwards if D_A increases); while a decrease in the parameters shifts E to the right (and upward). Fig. 2 shows the equilibrium curves for 'low' (solid line) and 'high' (broken line) values of the parameters, respectively.

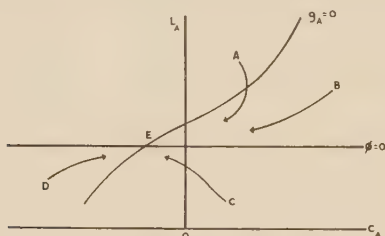


FIG. 1.—Possible Paths of the Systems
Arrows show possible paths (i.e., simultaneous changes in L_A and C_A) for various initial points, A, B, C and D . The system will be in equilibrium when it reaches E ; for then, $\phi = 0$ (see equation 2.1) and $g_A = 0$ (equation 2.2); hence, L_A and C_A will remain constant.

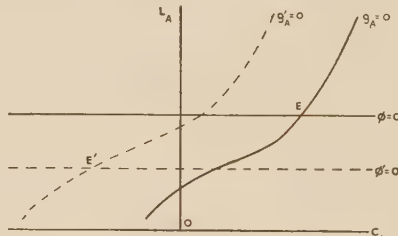


FIG. 2.—Positions of Equilibrium
 E and E' are equilibrium positions of the system corresponding to different values of the parameters appearing in equations (2.1) and (2.2).

For the particular values selected here, it may be seen that the equilibrium value of C_A is positive with low parameter values, but negative with high: that is, a confederate will be rejected by the group [$C_A(T) < 0$] if he persists in a very deviant opinion (D_A large), if the group is highly cohesive (C large) and if the item under discussion is highly relevant to the group's goal (R large). Under the opposite conditions, the confederate will not be rejected. These conclusions are all consistent with the data of Schachter's experiment [2, pp. 198 and 199, Tables III and VI].²

It will simplify matters if we regard the system as approximately linear, and measure the variables as deviations from the equilibrium values. That is, we treat the partial derivatives as constant, and take zero as the equilibrium value for all the variables. The system then becomes:

$$P_A = P_D D_A + P_U U_A. \quad (2.5)$$

$$\frac{dL_A}{dt} = \phi_D D_A + \phi_L L_A. \quad (2.6)$$

$$\frac{dC_A}{dt} = g_D D_A + g_U U_A + g_C C_A + g_L L_A. \quad (2.7)$$

$$U_A = U_C C_A + U_R R + U_C' C. \quad (2.8)$$

¹ For the methods employed here, see Lester R. Ford, *Differential Equations*, pp. 9-12, and A. A. Andronow and C. E. Chaikin, *Theory of Oscillations*, pp. 8-12, 182-193, and 203-208.

² It should be noted that the measures of C_A in Schachter's experiment did not determine the zero point (i.e., the line between acceptance of A and rejection). The data show that the direction of shift of E is that predicted by the model.

Four possible paths of the system, originating at points a, b, c, d , respectively, are shown in Fig. 3. Because we have taken the equilibrium point of the system as origin, a shift to the right of the initial point of the system is precisely equivalent to a shift to the left of $g_A = 0$ in the previous representation of the model. Hence a, b , and c may be regarded as three possible initial points, with the same initial values of L_A , but with progressively larger values of C , and/or R : (these larger values, as we have seen, shift g_A to the left relative to a fixed origin, or the initial point to the right relative to an origin at E).

From the shapes of the paths originating at a, c , and b , we conclude:

Case (1): If C and R are low enough, C_A will increase during the trial from its initial value towards the equilibrium value (Path a);

Case (2): If C and R are high enough, C_A will increase during the trial from its initial value toward the equilibrium value (Path c);

Case (3): For some intermediate range of values of C and R , C_A will first increase and then decline during the trial (Path b).

Now for a constant value of D_A , we see from (2.5) that the changes in P_A will be proportional to the changes in U_A , and from (2.8) that for constant R and C , the changes in U_A will be proportional to the changes in C_A . Hence the changes in P_A will be proportional to the changes in C_A . We conclude that the time shape of the curve $P_A(t)$ will be the same as that shown in Fig. 3 for $C_A(t)$.

It is in this way that we would interpret the data presented in Schachter's Table VII [2, p. 203] which measures P_A by the mean number of communications addressed to the 'deviate' during each of four ten-minute periods in the course of the experiment. For higher values of R and C in the Schachter experiment, $P_A(t)$ fell in Case 3, above. When C or R were at lower values, $P_A(t)$ had a time path corresponding to Case 1. Case 2 was not observed, suggesting that despite Schachter's laudable attempt to "reproduce the variables and phenomena under study with greater intensity in a purportedly 'real-life situation'" [2, p. 195], he had not succeeded in reaching high enough levels of C and R to produce the second case. However, he was successful in producing interesting qualitative differences among the time paths of $P_A(t)$ by using his four experimental combinations of high and low levels of C and R .

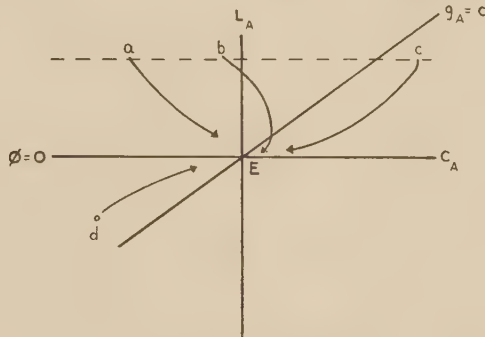


FIG. 3.—Shape of Time Path.
The arrows illustrate dependence of the shape of the time path upon the values of C and R .

III. COMPARISON OF DEVIATE-MEMBER AND AGGREGATE MODELS

As mentioned in an earlier paper [3], we have developed a model encompassing five propositions stated by Festinger [1] which deal with aggregate relations, i.e., are concerned with the members of a group taken as a whole. The model consists of the following mechanisms:

$$\frac{dD}{dt} = f[P(t), L(t), D(t)], \quad (3.1)$$

$$P(t) = P[D(t), U(t)], \quad (3.2)$$

$$L(t) = L[U(t)], \quad (3.3)$$

$$\frac{dC}{dt} = g[D(t), U(t), C(t)], \quad (3.4)$$

$$U(t) = U[C(t), R], \quad (3.5)$$

Mechanisms Involved in Group Pressures on Deviate-Members

where the variables are defined as follows:

$D(t)$: The perceived *discrepancy* of opinion on an issue among members of a group at time t ;

$P(t)$: Pressure upon members of the group to communicate with each other at time t ;

$L(t)$: Receptivity ('*Listening*') of members of the group to influence by communications from other members at time t ;

$C(t)$: *Cohesiveness* (identical with the corresponding variable in the deviate-member model).

$U(t)$: Pressure felt by the group to achieve *uniformity* of opinion, i.e. to reduce perceived discrepancy of opinion at time t ;

R : *Relevance* (identical with the corresponding parameter in the deviate-member model).

Now let us contrast this aggregate model with the deviate model. The variables in the two models are quite analogous; but the equations are somewhat different.

It will be noticed that the mechanism postulated in equation (1.1) is the same as in (3.2), but in the former case aggregated only over the subgroup excluding A . The mechanism in (1.3) is similar to that postulated in (3.4), but has been elaborated slightly to incorporate the effect of L_A —perceived receptivity of A to influence—upon C_A —the cohesiveness of the group with A . This mechanism would not produce essential differences in the behaviour of the system unless, as in the Schachter situation, there were forces tending toward subgroup formation.

The mechanism that makes the presence of L_A in (1.3) important is equation (1.2). This states that, when there is a discrepancy between A 's opinion and that of other members,

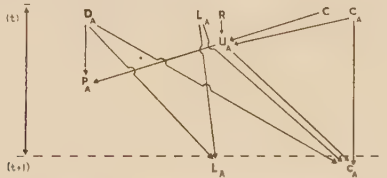


FIG. 4a.—Deviate-Member Model

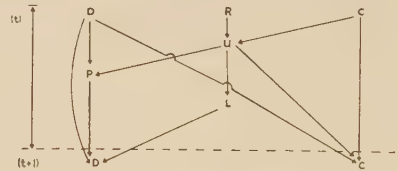


FIG. 4b.—Aggregate Model.

DIAGRAMS OF CAUSAL RELATIONS

Arrows are drawn to each dependent variable from each independent variable that influences it directly, as postulated in the equations of the system.

and as this discrepancy persists, the group will begin to *perceive* that A is unreceptive. This perception in turn, through equation (1.3), reduces the cohesiveness of the group with A . Equation (1.2) was not of importance in the aggregate model because there was no reason to postulate that the members would be highly unreceptive.

Finally, equation (1.4) represents a generalization of (3.5) to take account of the possibility of subgroup formation. In the deviate model, pressure toward uniformity with A depends *both* on the internal cohesiveness of the rest of the group (C) *and* on their cohesiveness with A (C_A), a distinction we did not have to make in the completely aggregative model.

By way of summary we give a diagram in Fig. 4a and 4b of the causal relations among the variables implied by the two systems of equations. For simplicity we assume change in variables to take place at discrete intervals rather than continuously. Thus, $L_A(t+1)$ is the value of L_A one time interval after t —similarly for C_A , D , and C .

The deviate-member and aggregate models may both be regarded as special cases of a more general implicit model in which aggregates do not appear, but only variables for individual members of the group. One special case (corresponding to our aggregate model) arises when the group behaves in an 'undifferentiated' manner, with similar mechanisms operating for all members. Another arises when the behaviour of one group member (or a small minority) becomes differentiated from the others, while the behaviour of those others can still be aggregated—our deviate-member model. When they are justified empirically it is advantageous to make these simplifying assumptions rather than deal with the more cumbersome general model. Similarly, in treating certain of the empirical studies with our aggregate model, we have ignored particular equations when the mechanisms they represented were not operating in the empirical data being examined. In formulating the deviate-model we added implicitly a mechanism of primary importance that needed to be only implicit in the aggregate model. This does not mean that we are explaining the several studies in terms of *different* theories, but rather that we can conceptualize a general theory wherein the particular phenomena under study

represent special cases. For the present it seems unprofitable to write the more general equations, which would consist of composite sets of equations analogous to those used in the two models, one set per group member. They would be complicated and difficult to handle, and at first glance do not seem capable of yielding fruitful derivations.

IV. COHESIVENESS AND REJECTION

Before concluding, we should like to comment on the cohesiveness variables, C_A and C , that appear in the two models. Cohesiveness is several times defined by Festinger and his colleagues as 'the resultant of all the forces acting on the members to remain in the group'. It plays the role of an intervening variable between (a) the various forces that increase or decrease the attractiveness of the group, and (b) the behaviours that result from these forces. Its usefulness as an intervening variable depends on whether, in fact, all the various types of positive and negative attraction to the group do produce the same kind of behaviour provided only that their net intensity is the same. The evidence on this is far from conclusive.

When C is defined as the 'resultant of all forces to remain in the group', the influence of the attractiveness of other groups is ignored. By interpreting C as the 'net attractiveness', it would seem that a wider array of phenomena could be encompassed in the aggregative model. Only one of Festinger's hypotheses considers the operation of the attractiveness of other groups:

Hypothesis 3c: The amount of change in opinion resulting from receiving a communication concerning 'item x ' will decrease with increase in the degree to which the opinions and attitudes involved are anchored in other group memberships or serve important need satisfying functions for the person.

With a 'net attractiveness' interpretation of $C(t)$, this hypothesis states a relation between ΔD and C via L and U , and is in fact identical with hypothesis 3b:

Hypothesis 3b: The amount of change in opinion resulting from receiving a communication will increase as the strength of the resultant force to remain in the group increases for the recipient.

That is, if C is the net attractiveness of the group, it can be increased either by increasing the forces to remain in the group or by decreasing the attractiveness of other groups. Hence the variable in 3b: 'strength of the resultant force to remain in the group' is the negative of the analogous variable in 3c: 'degree to which the opinions and attitudes involved are anchored in other group memberships'.

In this context, it may be well to call the reader's attention to our interpretation of the cohesiveness variables in the deviate-member model, wherein we regarded negative values of C_A as 'rejection'. This again is an effort to define the variables in the system more parsimoniously. The agreement of the deviate model with Schachter's data indicates that no new 'rejection' variable, apart from C_A is necessary to account for Schachter's results.

If one wanted to distinguish the attracting forces of other groups from the attracting forces of the group upon which attention is centred, two cohesiveness variables might be used, as C_1 and C_2 ; the 'net attractiveness' then being $C_1 - C_2 = C$, or cohesiveness.

There is another respect in which the cohesiveness variable is ambiguous. The consequences of low cohesiveness certainly must be expected to depend upon what alternatives a member sees to remaining in the whole group. There are at least four obvious possible reactions to a discordant group situation: (1) the member may consider the alternative of withdrawing from the group, (2) he may remain but become less sensitive to its attempts to control his opinions and behaviour, (3) he may consider the alternative of ejecting from the group one or more members who cause the discord, and (4) he may consider division of the group into subgroups. Hence, each member's perception of the distribution of group opinions, the possibilities of subgroup formation, his influence upon the group, and his opportunities for membership in other groups all affect the significance of cohesiveness. The predictions of our models all involved implicit assumptions with respect to these perceptions. The variable C would seem to be interpretable as the average intensity of the desire to maintain the group, but with the possible ejection of one or a few members. The variable C_A might be best interpreted as the average desire to retain a particular member of the group. No attempt was made to carry out these refinements. We do wish, however, to call attention to their importance in the further development of models of this kind.

Comment on Mathematization. The utility of mathematics in dealing with phenomena of the kind under discussion here is by no means universally accepted. What advantages do we think are gained by such an undertaking? We have tried, in examining Festinger's

Mechanisms Involved in Group Pressures on Deviate-Members

propositions, to illustrate concretely some of the advantages derivable from the construction of a formal mathematical model:

(a) Knowing more precisely what mechanisms or structural relations are being postulated, and sometimes calling attention to the need for further clarification of the operational meaning of definitions and statements;

(b) Discovering whether certain postulates can be derived from others, and hence can be eliminated as independent assumptions; whether additional postulates need to be added to make the system complete and the deductions rigorous; and whether there are inconsistencies among the postulates;

(c) Assisting in the discovery of inconsistencies between the empirical data and the theories used to explain them;

(d) Laying the basis for the further elaboration of theory, and to deductions from the postulates that suggest further empirical studies for verification;

(e) Aiding in handling complicated, simultaneous interrelations among a relatively large number of variables, with some reduction of the obscuring circumlocutions entailed by non-mathematical language.

In the long run, mathematics will be used in the social sciences to the extent that it provides a sufficiently powerful language of analysis and exposition to justify the time and effort required to use it. Social phenomena have proved sufficiently baffling; whether our inclinations are mathematical or literary, we cannot afford to prejudice what is the 'one best' methodology. Let methods, like substantive theories, prove by actual trial which is the more powerful, and under what conditions.

V. SUMMARY

1. In this paper we have examined certain hypotheses concerned with pressures toward uniformity upon the deviate-members of groups, and have endeavoured to develop them into an integrated, coherent system.

2. We have devised a model containing feed-back mechanisms to provide linkage among such variables as the perceived receptivity of the deviate-member and the desire of the group to retain him in the group.

3. Derivations from the model made it possible to check the system's congruence with empirical findings in an experimental situation. The check confirmed certain assumptions incorporated in the model, but showed also that many of the hypotheses set forth by Festinger and his associates simply were not being tested by the experiments. The mathematical derivations made it possible to predict time paths for certain of the variables and to check these predictions against the data.

4. A comparison was made of the deviate-member model with a homologous model developed for groups without a deviate. The two were found to be special cases of a more general system.

5. Two seemingly independent variables, cohesiveness and rejection, were found to be interpretable as merely opposite ends of a continuum. The new interpretation yielded derivations which were consistent with empirical data, and the consolidation helped increase the parsimony of the model.

APPENDIX. SYSTEMS OF DIFFERENTIAL EQUATIONS

The models with which we have been dealing attempt to explain a set of phenomena in terms of four or five mechanisms, each mechanism being represented by a differential or algebraic equation. Implicit in this method is the assumption that systems can actually be observed, in field or laboratory, whose behaviour is determined by these mechanisms and which are isolated from other systems to a sufficient degree of approximation.

Consider a system of n differential equations in n variables:

$$\frac{dx_i}{dt} = f_i(x_1, \dots, x_n) \quad (i = 1, \dots, n) \quad (1)$$

Arrange these equations in order of the 'rapidity' of the adjustment processes: (cf. below, p. 101). Let us suppose that we observe the whole system at intervals of τ minutes (or milliseconds, or years) and over a total span of T minutes (or milliseconds, or years). Now if the system starts at time zero

from a disequilibrium position, certain of the variables (the first k , say, which adjust most rapidly) will effectively reach their equilibrium values (within, say, ± 1 per cent.) within τ minutes, and hence before the next observation of the system. For these we shall have at all times (approximately):

$$f_i(x_1, \dots, x_n) = 0 \quad (i = 1, \dots, k) \quad (2)$$

which gives k algebraic equations to determine the first k variables in terms of the remaining $(n-k)$.

Among the remaining $(n-k)$ variables, some may have such slow adjustment processes that their values at time T are effectively the same (within, say, ± 1 per cent.) as at time 0. Suppose that these are the variables numbered from $(m+1)$ to n . Then the corresponding equations can be eliminated from (1), and the variables treated as constant parameters.

We have left, then, in addition to equations (2), a set of $(m-k)$ differential equations of the form (1), from which we can determine the time paths of the variables $x_{k+1}, \dots, x_m$. We see that by appropriate selection of the time units, τ and T , and of the initial position, $(x_1^0, \dots, x_n^0)$, of the system we can select out various subsets of the equations (1) for study.¹ Moreover, if we study the behaviour of the system as it moves from various initial positions, we can study the effect of various of the parameters (i.e., the variables $x_{m+1}, \dots, x_n$) upon it.

If the system of (1) is linear, we can make these statements more explicit. The general solution of the system is then:

$$x_i = \sum_{j=1}^n a_{ij} e^{\lambda_j t} \quad (3)$$

where the x 's are measured from their equilibrium values, the λ 's are complex numbers and the a 's are constant. The system is stable if and only if all the λ 's have negative real parts. The more negative the real part of λ_k the more rapidly does the term $a_{ik} e^{\lambda_k t}$ approach zero. If we now arrange the terms in order, so that the λ with largest real part comes first, and so on, those adjustment processes will be rapid for which all the a 's except the first few in (3) are very small, while those adjustment processes will be slow for which all the a 's except the last few in (3) are very small.²

¹ This is not, of course, the only way of conceptualizing the compartmentalization of a general system into independent subsystems. For a very important proposal along different lines see W. Ross Ashby, *Design for a Brain* (1952), particularly Chap. XI, XIV-XVIII, and XXIV. The whole problem is related to the question of causal ordering. See H. A. Simon, 'Causal ordering and identifiability'. In T. Koopmans (ed.), *Studies in Econometric Method*, Cowles Commission for Research in Economics, Monograph, XIV, 1953.

² In conversation, Dr. John W. Tukey has pointed out that this three-way classification of processes into 'rapid' (high frequency), 'slow' (low frequency), and 'intermediate' (intermediate frequency) is involved in every empirical analysis of time series. The over-all period of observation, T , places a lower limit on the frequencies that can be examined, and the interval of observation, τ , an upper limit. While Dr. Tukey's comments refer to the oscillatory components of time series, and ours to the exponential components the point involved is the same: any given set of observations only permits us to explore a subset of the mechanisms at work, and these are the mechanisms which are neither too slow relative to T , nor too fast relative to τ .

REFERENCES

1. Festinger, Leon (1950). 'Informal social communication.' *Psychological Review*, LVII, 271-282.
2. Schachter, Stanley (1951). 'Deviation, rejection, and communication.' *J. Abnor. Soc. Psychol.*, XLVI, 190-209.
3. Simon, H. A. and Guetzkow, H (1955). 'A model of short- and long-run mechanisms involved in pressures toward uniformity in groups.' *Psychol. Rev.*, LXII, 56-68.

TEST RELIABILITY ESTIMATED BY ANALYSIS OF VARIANCE¹

BY CYRIL BURT

Historical Note. The idea of determining the 'precision' of an estimated quantity from a series of repeated measurements is due originally to Gauss (*Theoria Combinationis Observationum*, 1809). It was introduced into psychology by Fechner who proposed a *Präzisionsmass im Gauss'schen Sinne*, viz., $h = 1/\sigma\sqrt{2}$ (*Elemente der Psychophysik*, 1859, pp. 103 f.). Clarke Wissler was apparently the first to suggest that 'the precision of a test may be estimated by correlating successive trials' (*Psychol. Mon.*, III, 1901, p. 60). The correlation so obtained has generally been known as a 'reliability coefficient'²: it seeks to answer the question—'if this test were repeated, how far would the results agree?' The traditional way of raising the question would have been to ask: 'if this test were repeated, how much variation would there be in the results?' The suggestion that the 'reliability' of tests and examination marks might be more effectively studied by analysing the variation—with the aid either of factor analysis or of what Fisher has termed the analysis of variance—was, I believe, first put forward in a Memorandum I was asked to prepare for the International Institute Examinations Enquiry³ in 1934; and a number of theoretical investigations with these newer techniques have since been published.⁴ Their adoption for practical purposes has, until the last few years, been comparatively rare.

Definitions. Both in this *Journal* and elsewhere several writers have recently complained that, 'in spite of the vast literature that is available, psychologists still use the term reliability in a loose and ambiguous fashion'.⁵ What kind of 'agreement' or 'variation' is to be regarded as a manifestation of 'reliability' or the reverse? We must therefore begin by defining a little more exactly what it is we are seeking to assess. The implicit notion is that of 'error', which is

¹ *Editorial Note.* This paper is intended to form one of a series of elementary 'expository notes' which the Journal Committee has suggested should be published. The formulation of the problem and the deduction of the various algebraic equations are therefore not so rigorous as would be desirable in an exposition designed for more advanced students. Points that might appear technical or even pedantic to the ordinary reader I have, so far as possible, relegated to footnotes. The greater part is taken from the roneo'd *Laboratory Notes on Reliability* (1938) and *Analysis of Variance* (1943) referred to by Mr. Mahmoud in the paper that follows. These notes in turn were amplifications of the Memoranda drawn up for the Examination Enquiry Committee in 1931: (see *Marks of Examiners*, 1936, and *The Marking of English Essays*, 1941).

From the correspondence that followed Mr. Mahmoud's note on the 'assessment of reliability by analysing variance' (this *Journal*, VII, 1954, p. 61) it appeared that many readers would welcome a more accessible account of the methods described in these publications; and the Editor therefore hopes that this paper may serve to explain their theoretical basis and that Mr. Mahmoud's may sufficiently illustrate their practical application.

² The substitution of the word 'reliability' arises from the fact that in psychology most of the early experimentalists received their training in Germany. At the turn of the last century, when German writers began to substitute Teutonic terms for French, *Zuverlässigkeit* was proposed as a synonym for *Präzision*. It was, for example, adopted by Krüger and Spearman to designate Wissler's 'test-retest correlation' (*Zeitsch. f. Psychol.*, XLIV, 1906, pp. 48). In his English writings Spearman used the ordinary English translation for *Zuverlässigkeit*, namely, 'reliability'. Cattell and Thorndike use the word 'reliability' in the wider sense, to cover all measurements of precision: (cf. E. L. Thorndike, *Mental and Social Measurements*, 1904, chap. X and *passim*).

³ In preparing the memorandum I was greatly aided by the work and advice of my colleagues in the Statistical Department of University College, particularly Professor R. A. Fisher, Dr. P. O. Johnson, and Dr. J. Neyman; and in revising the notes for teaching purposes I was much indebted to various research students, particularly to Miss M. Howard and Mr. R. W. B. Jackson.

⁴ Cf. more particularly P. O. Johnson, 'Tests of Certain Linear Hypotheses and their Application to Some Educational Problems', *Statistical Research Memoirs*, I (1936), 57-93, R. W. B. Jackson, 'The Reliability of Mental Tests', *Brit. J. Psychol.*, XXIX, 267-87, and G. A. Ferguson, *The Reliability of Mental Tests* (1941) and refs.

⁵ Cf. A. S. C. Ehrenberg, 'The Reliability of Repeated Measurements', *Bull. Brit. Psychol. Soc.* No. 23 (1954), 3 f., and this *Journal*, VI, 41 f., and A. Heim, *The Appraisal of Intelligence* (1954), 67 f.

Test Reliability Estimated by Analysis of Variance

itself a somewhat elusive concept. For statistical purposes a working definition may be derived as follows.¹

In psychological research, the investigator continually requires to measure as accurately as possible certain aspects or types of behaviour (which we will call 'traits') manifested by certain individuals. The most elementary observation that he can make, therefore, will consist in using (i) some experimental device, which we call a *test*, to elicit and assess the specified trait (ii) in some particular *individual* (iii) on some particular *occasion*. In any such experiment, the particular test, the particular person, and the particular occasion will each be specimens selected from a whole universe or population of such tests, persons, and occasions. We are thus concerned with variations in three distinct dimensions (see Fig. 1).

Accordingly, let x_{ijk} denote the observed measurement actually obtained for the i th individual with the j th test on the k th occasion ($i = 1, \dots, N$; $j = 1, \dots, n$; $k = 1, \dots, m$); let ξ_{ij} be the 'true' measurement, i.e., the measurement we should obtain if we possessed a perfectly efficient 'test'²; and let ϵ_{ijk} denote the error of the measurement actually obtained. Then, adopting algebraic addition as the simplest way of relating these quantities, we may write

$$x_{ijk} = \xi_{ij} + \epsilon_{ijk} \quad \text{OR} \quad \epsilon_{ijk} = x_{ijk} - \xi_{ij}, \quad \dots \quad (i)$$

and we may take the latter equation as a formal definition for the 'error' of a single observation.

This formulation involves two unknowns, but only one equation. To obtain equations that are soluble, the most obvious procedure would be (i) to secure measurements for the same person with the same test on a number of *different occasions*. In practice this may not always be feasible. Often it would be far more convenient (ii) to obtain measurements from the same person on the same occasion with a number of *different tests* (e.g., different problems or 'items' in a composite group test), or (iii) to obtain measurements on the same occasion with the same test from a number of *different persons*, or finally (iv) to use some combination of these various alternatives. We could then extract an estimate, not for the amount of *each* error (ϵ_i), but for the general tendency to error: e.g., with the first procedure

$$\bar{\epsilon}_{ij} = \frac{1}{m} \sum_{k=1}^{k=m} |x_{ijk} - \bar{x}_{ij}| \quad \text{OR} \quad s_e = \sqrt{\left\{ \frac{1}{m-1} \sum_{k=1}^{k=m} (x_{ijk} - \bar{x}_{ij})^2 \right\}},$$

where m as before denotes the number of occasions on which the test is applied.

Types of Error. Following Gauss and Fechner, psychologists commonly distinguish two main kinds of error: (i) systematic or 'constant' errors, which are more or less predictable and so avoidable; (ii) chance or 'random' errors, which are unpredictable and unavoidable. Here we shall be

¹ As Dr. Ehrenberg points out, most definitions involve a number of dubious assumptions (*loc. cit.*, p. 41): some of them indeed I myself believe to be unnecessary. However, Dr. Ehrenberg's criticisms of Professor Gulliksen's assumptions in *The Theory of Mental Tests* (1950) appear unduly severe: but cf. Dr. Guttman's review and Dr. Gulliksen's reply (*Psychometrika*, XVIII, 123-34).

² On the whole this simple definition of a 'true measurement' seems to me the best: I have phrased it so as to cover 'true measurements' that are purely hypothetical quantities as well as those (if any) which are actual quantities, e.g., Tom's innate general ability (which by hypothesis is unchanging) as well as, say, Tom's visual acuity or his efficiency in arithmetic. But the popular notion that this or that characteristic has in fact a 'true' measurement, if we could only find a perfect instrument for discovering it, is to my mind a highly dubious 'metaphysical' concept ('metaphysical' in the pejorative sense in which the word is used by the logical positivists). In psychology more particularly almost every specifiable trait is an unstable, fluctuating characteristic. Its measurement involves the determination of a region rather than a point. We can, if we like, define the particular point which, as it were, forms the centre of gravity about which the momentary values appear to vary. But *these internal variations do not form part of the 'unreliability' of our test or other metrical technique.*

To circumvent what I have called the metaphysical difficulty, it is tempting to fall back on some kind of operational definition: the 'true' measurement we may say is merely 'the limiting value towards which the mean of a number of measurements progressively tends as the number is indefinitely increased', or (in more modern terminology) 'the mathematical expectation over the whole universe of trials': this would

be equivalent to defining it by some such equation as $\xi = \mathcal{E}(x) = \frac{1}{m} \sum_{k=1}^{k=m} x_{ijk} \quad (m \rightarrow \infty)$. But there are pitfalls

in this definition. In particular, we shall be liable to include in the 'error' the variations due to changes in the trait itself as well as those due to the imperfections of the method of measurement. Hence I should prefer to treat this equation as an *inference* from the definition I have given above. For the general approach adopted in this section see F. N. David, *Probability Theory for Statistical Methods* (1949) chap. X; in accordance with the continental practice followed by Dr. David and others, I use $\mathcal{E}$ rather than E for such

operations as $\frac{1}{m} \sum_k$.

concerned almost exclusively with random errors. A random effect is usually described as one that results from a very large number of very small independent causes, about which nothing is known that could serve as a basis for prediction. I should prefer to follow Keynes, and define both 'randomness' and the related notion of 'independence' in terms of 'irrelevance'. From these definitions¹ several important corollaries can then readily be deduced. If the error e is a random variable in the sense defined, then as m increases,

$$(i) \mathcal{E}e_k \equiv \frac{1}{m} \sum_k^m e_k \rightarrow 0, \text{ and hence } (ii) \mathcal{E}x \equiv \bar{x} \rightarrow \xi;$$

$$(iii) r_{e_j e_{j'}} \equiv \mathcal{E}(e_{jk} e_{j'k}) \rightarrow 0; \text{ similarly, } (iv) r_{\xi e} \rightarrow 0;$$

and (v) with measurements of the kind with which the psychologist has to deal, the distribution of the errors will tend to the normal. With suitably defined conditions these corollaries will also hold, not only when the expectations refer to data obtained on a number of different occasions, but also when they refer to a number of different persons; and most of them can be subjected to an empirical check.

Definition of Reliability. (a) In Terms of Variation. Since we frequently need to compare the reliability of different tests, we cannot leave the measurements of error in units specific to each test or trait. The time-honoured device is to take as unit the standard deviation of the general population, or (by preference, since it is additive) the square of the standard deviation, i.e., the variance. This can be estimated by choosing a random sample of the population, and calculating

$$s_x^2 = \frac{1}{N-1} \sum_i^N (x_i - \bar{x})^2. \text{ On comparing the two variances it would then seem possible, on intuitive}$$

grounds, to infer that, when the variance of the measurements for a single individual (say the i th, $s_{e_i}^2$) becomes as large as the variance for the entire sample of different individuals (s_x^2), the test used will be of no practical value whatsoever: for the whole object of such a test is to *distinguish* the ability as measured for any given individual from the abilities of the rest. On the other hand, if the variance of the measurements for a single individual were zero, we should pronounce our method of measurement to be entirely free from error. The unreliability of the test will then be measured by the ratio of the individual variance to the total variance treated as unity. And for any given test j we may calculate the quantity

$$r_v = 1 - s_{e_j}^2 / s_x^2$$

which will evidently be 0 when $s_{e_i}^2 = s_x^2$, and 1 when $s_{e_i}^2 = 0$; and we may regard this as estimating on a scale of 0 to 1 the precision or reliability of the test. It will be seen that this approach treats reliability as an *invariant property of the test* (where 'test' means not merely the test-material but the entire process of measuring the specified trait for the specified population).

So far as the assumptions we have made hold good with the sample of persons actually measured, it will be a matter of indifference whether our sample of trials is obtained by repeated testing of one and the same typical individual or by testing a number of different individuals³; for, by corollary

¹ *Treatise of Probability*, pp. 55, 120, 291, 412; cf. also M. G. Kendall, *Advanced Theory of Statistics*, I, p. 171 f. and refs. p. 184, and David, *loc. cit. sup.* In the notation adopted by Keynes, X_1 is said to be irrelevant to X_2/h , if $X_2/X_1h = X_2/h$, i.e., if the probability of X_2 on the evidence X_1h is the same as its probability on the evidence h alone. Now let $C(X)$ mean 'the variable X is a member of the class C ', and let $\phi(X)$ mean ' X has the characteristic or the value ϕ '; then a is said to be a 'random' member of the class C , if the statement ' X is a ' is irrelevant to the probability $\phi(X)/C(X).h$, or (better perhaps) if $\phi(a)/C(a).h = \phi(b)/C(b).h$, where $C(b).h$ contains no information about b , except that it is also a member of the class C . The 'theorem of independence' states that if $X_2/X_1h = X_2/h$, then $X_1/X_2h = X_1/h$, and X_1/h and X_2/h are said to be independent. 'Statistical independence' is commonly defined in terms of statistical frequency functions: if the variation of X_1 is irrelevant to that of X_2 , i.e., if the distribution of X_2 is the same whether X_1 varies or not, then X_1 and X_2 are said to be independent; and in particular $f(X_1, X_2) = f_1(X_1)f_2(X_1)$ where $f(X)$ as usual denotes the 'frequency function' of X : (cf. Kendall, *loc. cit.*, p. 21, eqn. 1.22). Furthermore, $\mathcal{E}X_1X_2 = (\mathcal{E}X_1)(\mathcal{E}X_2)$ or $\mathcal{E}(X_1 - \mathcal{E}X_1)(X_2 - \mathcal{E}X_2) = 0$: in other words, the covariance is zero. This provides a useful test of independence: unless the covariance of X_1 with X_2 is statistically non-significant (and not always even then) we cannot justifiably assume that X_1 and X_2 are statistically independent.

² To avoid complicating the argument in the text I have omitted certain qualifications which a more rigorous statement would require; e.g., even supposing that the expectation of the sample mean is equal to the population mean (p. 104, n. 2), this would not of itself imply that the sample mean converges in probability to the population mean: still, under conditions which can nearly always be assumed in practice for data of the type here considered, it will converge.

³ At this point, I think, my treatment diverges from that adopted by Dr. Guttman ('Basis for Analysing Test-Retest Reliability', *Psychometrika*, X, 1945, pp. 255-82—an article which every student should study). Dr. Guttman restricts error variance to the variance of measurements for the i th individual, taken about the

Test Reliability Estimated by Analysis of Variance

(iv), there should be no relation between the size of the errors and the size of the true measurements. But in practice, if there *were* a possibility of some such relation, it would plainly be safer to draw our sample of errors from the whole range over which the test is to be applied. On these grounds, as well as those of ordinary convenience, it will be better to measure *all* the individuals at least twice, and then substitute the *mean* of the error variances in the numerator and the variance of the observed *means* for the different individuals in the denominator thus obtaining an alternative formula :

[illegible]

Now, even if the test had no diagnostic value, the means of the measurements for the different individuals would not be exactly the same. We should in fact expect their variance to be approximately equal to the error variance. Hence, when the variance of the observed means exceeds the error variance, we may reasonably attribute this excess to differences in the true measurements: in fact, it follows algebraically from eqn. (i) that, if (as we have assumed) the true measurements and the errors are uncorrelated, $\sigma_{\bar{x}}^2 = \sigma_{\varepsilon}^2 + \sigma_e^2$. Accordingly, we may estimate the variance of the 'true' measurements from the difference between the two. Thus, in practice, the formulae just reached are equivalent to saying that the reliability of the test can be defined as that proportion of the *total* variation observed which is attributable to the variation of the *true* measurements: i.e.,

$$rv = \frac{\frac{s_x^2 - s_e^2}{2}}{s_r} = \frac{\hat{s}_\xi^2}{\frac{2}{s_r}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (iii)$$

where the circumflex accent denotes that the variance of ξ is merely an estimate.

(b) *In Terms of Agreement.* If the differences between the measurements obtained for one and the same quantity provide an indication of their unreliability, it is natural to treat their agreement as an indication of reliability. At its simplest this agreement can be assessed by computing the correlation between the two series of measurements obtained with identical forms of a given test (j and j'), applied under identical conditions, namely,

$$r_{xx'} = \frac{\sum_i x_{ij} x_{ij'}}{\sqrt{(\sum_i x_{ij}^2) (\sum_i x_{ij'}^2)}} \quad (i = 1, 2, \dots, N).$$

Now $\sum x_{ij} x_{ij'} = \sum (\xi_i + \epsilon_{ij})(\xi_i + \epsilon_{ij'}) = \sum \xi_i^2 + \sum \xi_i \epsilon_{ij} + \sum \xi_i \epsilon_{ij'} + \sum \epsilon_{ij} \epsilon_{ij'}$, and this (by corollaries ii and iii above) tends to $\sum \xi_i^2$ as N increases. Similarly $\sum x_{ij}^2 \rightarrow \sum \xi_i^2 + \sum \epsilon_{ij}^2$ and $\sum x_{ij'}^2 \rightarrow \sum \xi_i^2 + \sum \epsilon_{ij'}^2$. Moreover, since by hypothesis the two forms of the test are identical, $\sum \epsilon_{ij} = \sum \epsilon_{ij'} = \sum \epsilon_i^2$. Hence, under the conditions assumed,¹ we may write

$$r_{xx'} = \frac{\sum \xi_i^2}{\sum \xi_i^2 + \sum \epsilon_j^2} = \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_e^2} \quad . \quad . \quad . \quad . \quad (iv)$$

or, substituting estimates,

$$= \frac{s_x^2 - s_e^2}{s_x^2} = 1 - \frac{s_e^2}{s_x^2} = r_v. \quad (v)$$

In passing we may note that, since

$$r_{xx'} \rightarrow \frac{\sigma_{\xi}^2}{\sigma_x^2} \quad \text{and} \quad \frac{\sigma_{\xi}}{\sigma_x} = \frac{\sigma_{\xi}^2}{\sigma_{\xi}\sigma_x} = \frac{\sum \xi x}{\sqrt{(\sum \xi^2 \sum x^2)}} = r_{\xi x},$$

we may take $\sqrt{r_{xx}}$ as an estimate of $r_{\xi X}$, the 'index of reliability' (Kelley, *Statistical Method*, 1923, p. 206, eqn. 160). Hence, unless the 'criterion' (i.e., the independent estimates obtained for the

mathematical expectation and over 'an indefinitely large *universe of trials*'. This gets rid of many of the assumptions otherwise required. Now, although an economy of assumptions is desirable in *theoretical* work, in *practical* work it seems in general wiser to retain them, and so far as possible, to verify to what extent they are fulfilled.

¹ For eqn. iv to be valid a less stringent requirement would serve. As the proof implies, the necessary and sufficient condition is merely that the two columns of measurements to be correlated are homoscedastic.

trait) is itself as unreliable as the test, it is quite erroneous to state, as is so often done, that 'the validity of a test can never exceed its reliability'.¹

Equation iv expresses in its simplest form a mode of approach first suggested by Galton, namely, an analysis of variation into independent components or 'factors': if the total variance, σ_x^2 , can be divided into two independent and additive portions, one due to the factor *common* to both variables (σ_x^2) and the other due to the factors *specific* to either variable (σ_e^2), then the amount of agreement between the two sets of measurements can be expressed by the ratio of the common factor variance to the total variance.²

Owing to the fact that the non-statistical psychologist always finds it easier to think in terms of correlations, he has, as a rule, nearly always preferred to compute a kind of 'self-correlation' by applying the familiar product-moment formula. But the expression we have reached (eqn. v) shows that the same value can be got by adopting an analysis of variance instead of the more usual procedure. Such an analysis will have several obvious advantages: it will enable us to check the underlying assumptions, to obtain the same coefficient by a more direct calculation, to construct modified forms of the coefficient where required, to adapt the concept to cases where the test has been applied only once or generalize it to cases where the test has been applied more than twice, and to extend the treatment to investigations in which all three sources of variation—persons, tests, and trials—and even others as well, can be compared simultaneously.

I. THE RELIABILITY OF A SIMPLE TEST

Illustrative Example. The procedure proposed can be most simply explained if we start with an actual example. Let us suppose that, by means of a booklet comprising five sets of questions, we have measured the intelligence of six pupils, chosen at random, and have applied the test on two successive occasions. The detailed marks³ obtained at each of the trials are set out in Table I. The two rows of figures near the bottom of the table, which give the averages for the five subtests, will form the 'final marks' for the first and second trial respectively. The reader may, if he likes, think of the figures as 'intelligence quotients'. To deal with the problem in its simplest form first, we shall for the moment ignore the data for the constituent subtests, and merely ask what is the reliability of the composite test, judged by the 'final marks'.

If identical tests, or two equivalent (that is, interchangeable) forms of the same test, have been used for the two trials, we may expect that both the means and the standard deviations will be approximately equal on each occasion. Since the occasions are themselves samples, we cannot expect *exact* equality; but what small or non-significant discrepancies there may be can be treated as an incidental result of the unreliability. Accordingly, the best estimate of the mean for the entire sample will be the mean of the two averages obtained for the whole group on the two occasions (101.1). Similarly the best estimate for the standard deviation will be derived from the average of the two square-sums taken about this mean. The coefficient of correlation obtained from estimates so computed is termed the intra-class correlation; and, as I have argued elsewhere, it appears to be the most suitable form to adopt for this purpose. It has the merit of being based on $2N-1$ instead of $N-1$ degrees of freedom, and thus provides a more accurate estimate than the ordinary product moment coefficient. On this basis, therefore, the formula will be (R. A. Fisher, *Statistical Methods for Research Workers*, 1935, p. 199)

$$r_{int} = \frac{\sum \{(x - \bar{x})(x' - \bar{x})\}}{\frac{1}{2} \{ \sum (x - \bar{x})^2 + \sum (x' - \bar{x})^2 \}} \quad \text{. (vi)}$$

¹ Several writers have argued that, if their tests have been shown to be valid, 'they must therefore be reliable' (cf. H. J. Eysenck, this *Journal*, VI, 1953, pp. 44). Some of Dr. Eysenck's tests, for example, have theoretical validities of over 0.70—admittedly a high figure for tests of personality; but a test with a theoretical validity of 0.70 might have a reliability of only $0.70^2 = 0.49$, which is not a satisfactory value for a reliability coefficient.

² It is instructive to compare the ratio here reached with the expression for 'information communicated' in arguments from information theory. To take the simplest case, let the 'true message' to be communicated be denoted by ξ and the obscuring 'noise' by ϵ , and let the message as received be denoted by

$x = \xi + \epsilon$: then, as Wiener proves, the information gained can be measured by $\frac{1}{2} \log_2 \frac{\sigma_\xi^2 + \sigma_\epsilon^2}{\sigma_\epsilon^2}$. The formula

can be generalized to meet the more complex cases where a multiplicity of variables and common factors are involved (cf. N. Wiener, *Cybernetics*, 1948, and Burt, this *Journal*, IV, 1951, pp. 193 f.). In my fuller *Notes* I endeavoured to show that a more satisfactory derivation of the formulae for reliability could be obtained if the basic definitions and postulates were expressed in terms of information theory. However, this line of approach seemed too technical for adoption here.

³ The detailed figures are taken from K. Mather, *Statistical Analysis* (1943), p. 73, Table 12. The reason for the choice will be explained in a moment (p. 111). To save space the tables give values to one decimal place only: the calculations are based on two or more.

Test Reliability Estimated by Analysis of Variance

where x denotes the first set of N measurements and x' the second set, and $\bar{x}$ denotes the mean of the entire series. The formula, it will be seen, departs from the more familiar version by substituting the arithmetic mean of the two variances for the geometric.

Applying it to the average marks obtained by the several persons at the two successive trials (the 'final marks' shown in the last two lines but one of Table I), we obtain

$$r_{int} = \frac{1052.85}{\frac{1}{2}(3945.81 + 2436.84)} = 0.3299.$$

With the same data the ordinary inter-class correlation would be $r = 0.5288$. The discrepancy, here unusually large, is due mainly to the rather big difference between the means for the two trials.

The Coefficient as a Ratio of Variances. The same value can be reached by calculating the three appropriate variances from the square-sums (i) of the observed values, (ii) of the means for the several persons, and (iii) of the deviations about these means.

TABLE I. UNSTANDARDIZED MEASUREMENTS

Test	Occasion	John	James	Hugh	George	Ralph	Henry	Average
M	1	81.0	146.6	82.3	119.8	98.9	86.9	102.58
M	2	80.7	100.4	103.1	98.9	66.4	67.7	86.20
M	Av.	80.8	123.5	92.7	109.3	82.6	77.3	94.39
S	1	105.4	142.0	77.3	121.4	89.0	77.1	102.03
S	2	82.3	115.5	105.1	61.9	49.9	66.7	80.23
S	Av.	93.8	128.7	91.2	91.6	69.4	71.9	91.13
V	1	119.7	150.7	78.4	124.0	69.1	78.9	103.47
V	2	80.4	112.2	116.5	96.2	96.7	67.4	94.90
V	Av.	100.0	131.4	97.4	110.1	82.9	73.1	99.19
T	1	109.7	191.5	131.3	140.8	89.3	101.8	127.40
T	2	87.2	147.7	139.9	125.5	61.9	91.8	109.00
T	Av.	98.4	169.6	135.6	133.1	75.6	96.8	118.20
P	1	98.3	145.7	89.6	124.8	104.1	96.0	109.75
P	2	84.2	108.1	129.6	75.7	80.3	94.1	95.33
P	Av.	91.2	126.9	109.6	100.2	92.2	95.0	102.54
Final Marks	Av. 1	102.8	155.3	91.8	126.2	90.1	88.1	109.05
Final Marks	Av. 2	83.0	116.8	118.8	91.6	71.0	77.5	93.13
Final Av.		92.4	136.0	105.3	108.9	80.6	82.8	101.09

Expressing the average marks (near bottom of Table I) as deviations about the final mean, we obtain the following figures :

							Square-Sums	Divisor	Variance	
(i)	{ Av. 1	+ 1.7,	+54.2,	- 9.3,	+25.1,	-11.0,	-13.0}	6382.7	12	531.9
	{ Av. 2	-18.1,	+15.7,	+17.7,	- 9.5,	-30.1,	-23.6}			
(ii)	Means	- 8.2,	+34.9,	+ 4.2,	+ 7.8,	-20.5,	-18.3	2122.1	6	353.7

And the deviations from these means will be

(iii)	{ for Av. 1	+ 9.9,	+19.3,	-13.5,	+17.3,	+ 9.5,	+ 5.3}	2138.5	12	178.2
	{ for Av. 2	- 9.9,	-19.3,	+13.5,	-17.3,	- 9.5,	- 5.3}			

Dividing the square-sums by the number of items summed, we obtain for the total variance 531.9, for the variance of the means 353.7, for the variance of the individual deviations 178.2. Had the test no diagnostic value, we should expect the observed means to show a random variance, similar

to the variance of the individual deviations. We may therefore take $353.7 - 178.2 = 175.5$ as our estimate of the variance of the 'true' measurements; and thus obtain for our estimate of the reliability (in accordance with equation (iv))

$$r = \frac{\sigma_{\xi}^2}{\sigma_{\xi}^2 + \sigma_{\epsilon}^2} = \frac{175.5}{531.9} = 0.3299,$$

the same figure as we reached by the method of correlation.

Analysis of Pooled Variance. For practical computation the foregoing procedure is needlessly cumbersome. By far the quickest method will be to compute, direct from the observed values in the usual way, the square-sums needed for a simple analysis of variance.¹ This will at the same time enable us to test the statistical significance of the coefficient so computed. The complete analysis is shown in Table II. Note that, since the means are, as it were, latent in *both* sets of measurements (that for the first trial and that for the second), the sum of square-sums 'between means' must be entered as 2×2122.1 .

TABLE II. ANALYSIS OF VARIANCE

for Two Applications of a Composite Test

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio	Probability
Between means for persons	$Q_p = 4244.2$	$(N-1) = 5$	$V_p = 848.8$	2.38	<0.05
Within measurements for persons	$Q_r = 2138.5$	$N(m-1) = 6$	$V_r = 356.4$		
Sum	$Q_s = 6382.7$	$(Nm-1) = 11$	$V_s = 1205.2$		

The value for the intraclass correlation can now be obtained from the ratio of the *difference* of the two component square-sums to their *sum*.² We have in fact

$$(\text{biased}) r_{int} = \frac{Q_p - Q_r}{Q_p + Q_r} = \frac{4244.2 - 2138.5}{4244.2 + 2138.5} = \frac{2105.7}{6382.7} = 0.3299.$$

Unbiased Estimate. In point of fact, however, the formulae so far given for the required correlation yield an estimate which is biased,³ and in general too low. The reader, indeed, may have noticed that a moment ago, to obtain the three variances, I divided by 'the number of items summed'. With Fisher's method of carrying out analysis of variance, we obtain the 'mean squares' by dividing by the 'degrees of freedom'. And, in accordance with this principle, we can correct for the bias by using the difference and sum of the estimated mean-square in place of the difference and sum of the square-sums. We then have

$$(\text{unbiased}) r_{int} = \frac{V_p - V_r}{V_p + V_r} = \frac{848.8 - 356.4}{848.8 + 356.4} = \frac{492.4}{1205.2} = 0.4086.$$

The correction is here rather large; with bigger samples it is usually much smaller.

To test the significance of the value thus found, we calculate a variance-ratio in the usual way (2.38); and it will be seen that, if we had no other details about the results of our test, we should be led to infer that the differences between the means for the various persons were statistically non-significant. This inference, however, depends on the assumption that the 'mean square within persons', as given in the table, provides the best available estimate for the error of measurement. But in the original table of detailed marks we have far fuller data for making such an estimate which we have not yet used; and this further information, as we shall shortly discover, suggests that much of what is here included in the estimate of 'error' is really of the nature of 'constant error' rather than of 'random error', and consists largely of predictable variations due to peculiarities of the sub-tests and of the different occasions. Let us therefore now consider how far these irrelevant variations can be isolated or allowed for by undertaking a complete analysis of all the detailed marks.

¹ If, however, the observed values have already been reduced to deviations about the final mean, it will be quickest just to square (a) their sums and (b) their differences, and then sum the squares: this gives us 8288.4 and 4277.0—exactly twice the square-sum calculated in the usual way, i.e., $2Q_p$ and $2Q_r$, which we then insert in a formula analogous to that given in the next paragraph.

² See *Factors of the Mind*, p. 275, eqn. i, putting $k = m = 2$.

³ See Fisher, *loc. cit.*, p. 213, and *Factors of the Mind*, p. 275, eqn. ii.

II. THE RELIABILITY OF A COMPOSITE TEST

A. UNSTANDARDIZED MARKS

The Three-Way Analysis of Variance. When a composite test or battery has been applied to a group of testees on two or more occasions, then, as we have seen, we are really faced with three¹ main sources of variation: (1) the differences between the *persons* tested, (2) the differences between the *tests* employed, and (3) the differences in the conditions obtaining on various *occasions* when the tests are applied. Moreover, each of these main causes may *interact* with one another, so that, in all,

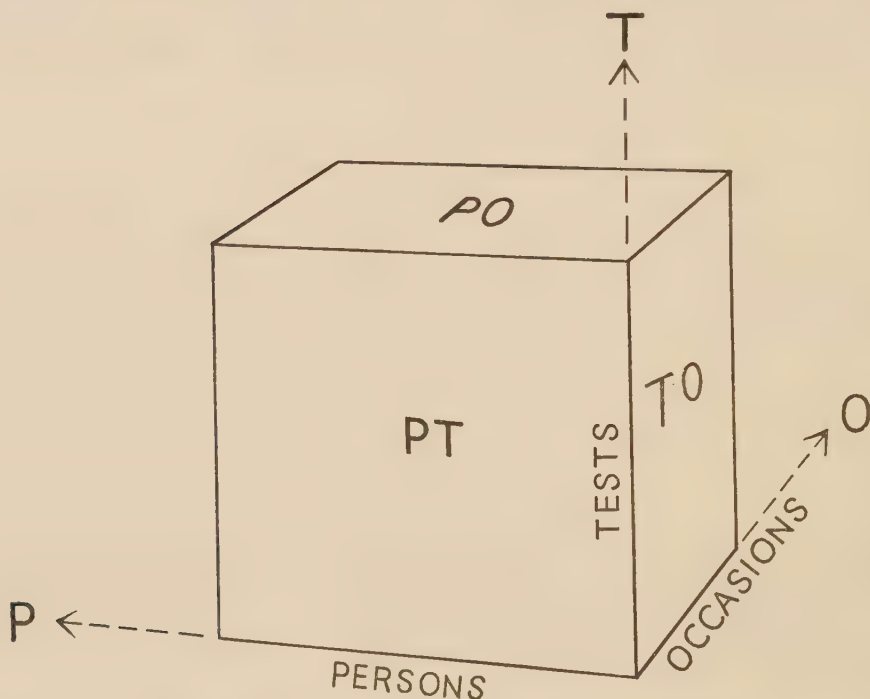


FIG. 1—Analysis of Variance with Three-Way Classification.
Main Dimensions of Variation and Interactions.

the complete analysis will include seven possible sources of variation, namely, three main sources, three first-order interactions, and one second-order interaction which will be used as an estimate of error.²

¹ In investigations into 'marks of examiners' there are frequently *four* main sources of variation: viz. the differences between (i) the question-papers or subjects set (e.g., different topics chosen for an English essay), (ii) the persons examined, (iii) the persons marking the scripts, (iv) the different occasions on which the examination is taken or the scripts marked. As investigations for the International Institute Examinations Enquiry Council showed, an adequate research would involve a specially planned experiment, adopting the method of the 'Latin Square'. For an illustration of this further method, cf. C. Burt and R. B. Lewis, *Brit. J. Educ. Psychol.*, XVI, 1946, pp. 116-32: the investigation is reproduced by H. M. Walker and J. Lev in the introduction to their chapter on 'Analysis of Variance with Two or More Variables of Classification' (*Statistical Inference*, 1953, pp. 348-86) to which the research student may refer for further guidance.

² For purposes of class-instruction I find it useful to illustrate these various possibilities by geometrical diagrams or models similar to those of Stern. In this case it will be a rectangular parallelepiped: the 3-dimensional solid represents the sample; the three axes of co-ordinates the three main sources of variation (the indefinite 'universes' of persons, tests, and occasions); the three coordinate planes the 1st-order interactions; and the volume the 2nd-order interactions or 'error'. Cf. Figure 1.

Our first task therefore is to analyse the total variance (or rather the total square-sum) into the contributions resulting from these various sources. The procedure in this part of our work will be identical with that described in most textbooks of statistics.¹ To save repeating the algebraic proofs involved and the detailed working instructions, the example I have chosen (Table I) consists of a set of data already fully analysed in what is perhaps the most illuminating discussion of the subject for the student of psychology. The student who wishes to follow the actual computations step by step will find them fully set out in the reference given.² The final results are shown in Table III.

TABLE III. ANALYSIS OF VARIANCE: UNSTANDARDIZED MEASUREMENTS

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio	Probability
<i>Main Effects</i>					
P. Persons	(Q_p) 21220.9	5	4244.2	30.5	<0.001
T. Tests	(Q_t) 5310.0	4	1327.5	9.5	<0.001
O. Occasions	(Q_o) 3798.5	1	3798.5	27.3	<0.001
<i>First Order Interactions</i>					
PT. Persons and Tests	(Q_{pt}) 4433.0	20	221.7	1.6	0.200-0.100
PO. Persons and Occasions	(Q_{po}) 6893.9	5	1378.8	9.9	<0.001
TO. Tests and Occasions	(Q_{to}) 291.8	4	73.0	—	—
<i>Second Order Interaction</i>					
PTO. Residual	(Q_{pto}) 2784.2	20	139.2	—	—
Sum (Q_s)	44732.3	59			

Main Sources of Variation. There are three preliminary problems to be considered: what evidence do the data afford for differences (i) in the general ability of the persons as tested by the battery as a whole, (ii) in the severity or difficulty of the several subtests, and (iii) in the conditions under which the tests are applied on this occasion or that? The answers are clear. The effects of all three main sources of variation are fully significant.

Of the three variance-ratios, that due to differences between the 'persons' tested is the largest, and now proves to be fully significant. These of course are the differences which our tests were designed to measure; and it is therefore encouraging to find that, even with so small a sample, the investigation yields results that are statistically significant.

The variance ratio due to differences between the two occasions is almost as large. With actual data we might infer that the difference was due to the fact that the second battery is slightly harder, or possibly to some unfavourable circumstance affecting the conditions of the second trial or again to some change in the pupils themselves: had the second mean been higher, we should doubtless have argued that familiarity or practice had improved their general performance; as it is, we might wonder whether the loss of novelty has produced a decline in interest and keenness.

The variance-ratio due to differences between the subtests is far smaller. Here the differences might be due either to differences in their difficulty or to differences in the method of marking. In an actual research such results would indicate that the tests were still capable of considerable improvement.

The Interactions. The interaction between persons and tests can hardly be regarded as significant; but the interaction between persons and occasions is fully significant; while in the case of the interaction between tests and occasions the variance turns out to be smaller than the residual variance and therefore entirely devoid of significance.

Students of psychology usually find the term 'interaction' obscure. Let us therefore glance in passing at the detailed figures which it here denotes. Consider first the interaction between persons and tests. To compute it we begin by averaging the pair of marks (one for each trial) obtained by

¹ E.g., M. G. Kendall, *The Advanced Theory of Statistics* (1946), II, pp. 187 f., K. Mather, *op. cit.*, pp. 72 f., or H. M. Walker and J. Lev, *op. cit.*, pp. 363 f.

² K. Mather, *Statistical Analysis in Biology*, with a Foreword by R. A. Fisher (1943), pp. 72-79 (Example 7). Mather's data were in fact agricultural: they might be described as the results of testing the crop-fertility of 6 different soils by using 5 different kinds of seed, sown on two successive occasions. But to make the procedure more readily intelligible to the student of psychology, I have imagined (as stated in the preceding section) that they represent the results of 5 mental tests. A simple illustration of a three-way analysis of variance with *actual* tests was given in an earlier paper in this *Journal* (I, p. 10, Table IV).

Test Reliability Estimated by Analysis of Variance

each boy in each of the tests. As there are 6 boys and 5 tests, we get 30 averages. We then reduce these averages to deviation form by columns and by rows: in other words, we subtract first the means for the several tests and then the means for the several persons (or vice versa). The final result is the 'doubly-centred measurement matrix' set out in Table IV. It will be seen that every row and every column adds up to zero. On squaring the figures and taking the sum, we find that the square-sum for these 30 values is 2216.5. But the original mark sheet contained 60 values; and there each boy's average mark for each test is incorporated in the initial data twice over, once for the first trial and once for the second. In fact Table IV only represents half the interaction, namely, that for the first trial: the other half, that for the second, would be exactly the same. Over the whole table, therefore, the interaction between tests and boys contributes $2 \times 2216.5 = 4433.0$ to the total square-sum; and this, it will be seen, is the value inserted in Table III above.¹

TABLE IV. INTERACTION BETWEEN PERSONS AND TESTS

Person	John	James	Hugh	George	Ralph	Henry	Total
Test M	- 5.34	- 5.84	- 5.91	7.15	8.79	1.15	0.00
S	10.92	2.67	- 4.16	- 7.30	- 1.15	- 0.98	0.00
V	9.06	- 2.69	- 5.95	3.11	4.25	- 7.78	0.00
T	- 11.55	16.45	13.18	7.14	- 22.07	- 3.15	0.00
P	- 3.09	- 10.59	2.84	- 10.10	10.18	10.76	0.00
Total	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Sum of squares = 2216.5; $2 \times 2216.5 = 4433.0$.

What does such a table mean? Inspection of the figures suggests that both tests and boys tend to fall into distinguishable types. James, Hugh, and George appear to do better than the rest at the T-test, and worse at the P- and M-tests; John, Ralph, and Henry, on the other hand, tend to do worse at the T- and S-tests, and better at the P- and M-tests: there are, however, marked exceptions. This doubly-centred matrix of measurements is in fact equivalent to the table of residuals from which, in an ordinary factor analysis, the bipolar factors would be derived after the general factor had been removed (though in practice such factors would be calculated from the residual correlations, not from the residual test measurements). Indeed the only way of deciding how precisely the boys or the tests should be classified would be to carry out the appropriate factorization. A doubly centred matrix like that we have just computed must have a rank of either $(n-1)$ or $(N-1)$, whichever is smaller. Here, therefore, had the sample of persons been large enough to justify such a procedure, we could have extracted as many as *four* bipolar factors.²

TABLE V. INTERACTION BETWEEN PERSONS AND OCCASIONS

Person	John	James	Hugh	George	Ralph	Henry	Total
1st Trial	1.97	11.33	- 21.50	9.30	1.56	- 2.66	0.00
2nd Trial	- 1.97	- 11.33	21.50	- 9.30	- 1.56	2.66	0.00
Total	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Sum of squares = 1378.8; $5 \times 1378.8 = 6893.9$.

The interaction between persons and occasions is fully significant. The detailed figures are shown in Table V. They suggest the possibility that experience with the tests in the first trial affects

¹ The calculation here described would in practice be a very cumbersome way of computing the square sum for the interactions; it is much quicker to subtract the square-sums for the tests and persons at the end of the computation (see Mather, *loc. cit.*, p. 75 and Table 14A).

² I have discussed this particular interaction at some length because, as Mr. Mahmoud remarks, my previous example (this *Journal*, I, p. 23) has apparently led certain readers to suppose that only *one* elementary factor was needed to explain each interaction. In that earlier illustration no more happened to be needed, because, to simplify the exposition, I selected an inquiry in which only two tests were applied on each of two occasions; hence all the interaction matrices there had a rank of one. For a detailed factor-analysis of a doubly-centred measurement matrix, such as that illustrated above, but obtained in a psychological research, see Burt, 'The Analysis of Temperament', *Brit. J. Med. Psychol.*, XV, 1938, pp. 166 f., esp. Table I.

different children differently in the second. Hugh, for example, seems to profit by it; while James and George apparently find the second trial less interesting, and consequently show a decline. Here we are dealing with only two occasions; hence the interaction can be expressed by a single factor only.

The interaction between tests and occasions is quite devoid of significance. Had it been significant, we might tentatively have inferred that the abilities tested by the different subtests were affected differently by the passage of time: in other words, the results would have been suggestive of 'function fluctuation'. (Any decisive proof of such changes would need a more elaborate experimental design: see below.)

It will be instructive to compare the figures for the more detailed analysis (Table III) with those for the previous analysis (Table II). In the previous analysis the five subtests were pooled (this abolished Q_t and Q_{to}); and the difference between the means for the two trials was provisionally taken to be non-significant (hence Q_o and Q_{po} were combined to yield Q_r). And whereas that analysis was based on two sets of figures only for each person, namely, his average mark at either trial, the present analysis is based on five sets of figures for each trial—ten sets in all. This has the effect of multiplying the former square-sums by 5. The relations between the two are shown in Table VI.

TABLE VI. COMPARISON OF ANALYSES

Source of Variation		Square Sums First Analysis		Square Sums Second Analysis	
P.	Persons	(Q_p)	$4244.2 \times 5 = 21220.9$		21220.9
O.	Occasions				3798.5
PO.	Interaction	(Q_r)	$2138.5 \times 5 = 10692.4$	{ (Q_o) (Q_{po})	6893.9
Sum		(Q_s)	$6382.7 \times 5 = 31913.3$	(Q_s)	31913.3

(a) *Marks Corrected for Difference between Means for Trials.* Since the difference between the means for the two trials has proved fully significant, we are now justified in regarding it as a source of constant error rather than of random error. It can be eliminated by converting the pooled marks into deviations about the mean for each year. The inter-class correlation remains as before; but the intra-class correlation now comes into closer agreement with it. Using the same formulae we have

$$r_{int} = \frac{1432.7}{\frac{1}{2}(3566.0 + 2057.0)} = 0.5096,$$

$$\text{or } \frac{(21220.9 - 6893.9) \div 5}{(21220.9 + 6893.9) \div 5} = 0.5096.$$

(b) *Marks Corrected for Differences in Test Scales.* The more detailed analysis has revealed peculiarities in the several subtests, resulting from differences either in difficulty or in severity of marking: these can be eliminated by converting the marks to deviation-form. There are, too, obvious differences in the range or spread of the marks, which must introduce an arbitrary weighting when the raw marks are summed as they stand: these can be abolished by reducing the deviations to standard measure.

B. STANDARDIZED MARKS

Correlations. To save space I shall not here reproduce the amended version of Table I which this standardization provides.¹ The changes, however, will inevitably modify the means or 'final marks' for the several boys. Hence the reliability coefficient must be calculated afresh. But there is yet another procedure for calculating such coefficients, which is highly instructive, and may therefore be briefly considered at this point.

In constructing a new battery, an experienced investigator will almost always calculate and examine, not only the correlation between the final marks, but also the correlations between the different subtests. If the marks have been reduced to standard measure, these can be obtained at once by simply computing the product-sums, and dividing by N . When the tests have been applied on two or more occasions, it is convenient to tabulate the coefficients so obtained in accordance

¹ Any reader who wishes to use this example for teaching purposes can obtain from the editor a copy of the detailed marks in standard measure and of the detailed intercorrelations.

Test Reliability Estimated by Analysis of Variance

with the scheme originally suggested for purposes of a group factor analysis,¹ arranging all the inter-correlations for the same day in square submatrices astride the leading diagonal. When there have been only two such applications, the entire correlation matrix will be quartered into four submatrices. The coefficients for the first day will then be placed in the N.W. quarter, since these will presumably have a group factor of their own; those for the second day in the S.E. quarter, since they will contain a different group factor; while the cross-correlations will be placed in the N.E. and S.W. quadrants, since they will involve neither group factor, but only the common factor (and perhaps specifics in the diagonal). The next step is to sum values in each of the four matrices; and, if a group factor analysis is to be carried out, the subtotals for the rows and columns will be required as well. We can then calculate what Pearson would have called the 'bimultiple correlation' between the two sets of tests, with or without differential weighting. The use of differential weights we can defer until we consider the application of factor analysis. Mr. Mahmoud gives an excellent illustration of the procedure in the paper that follows (cf. Table II, pp. 121 f.). To save space, therefore, I shall here omit the detailed table of correlations and give only the final results. As will be seen from Mr. Mahmoud's example, once the correlation table has been constructed, we have merely to take the sum of the figures in the N.E. (or S.W.) quarter for the numerator, and the mean of the sums in the N.W. and the S.E. quarters for the denominator. With the present data, if we take the geometric mean of the two last sums, we obtain the ordinary intercorrelation

$$r = \frac{9.8041}{\sqrt{(22.0912 \times 18.9680)}} = 0.4789;$$

if we take the arithmetic mean, we obtain the intra-class correlation

$$r_{int} = \frac{9.8041}{\frac{1}{2}(22.0912 + 18.9680)} = 0.4776.$$

The difference is now almost negligible.

Analysis of Variance. We shall obtain exactly the same figure for the intra-class correlation, if we carry out a detailed analysis of variance along the same lines as before. Our method of standardization has completely eliminated the variance between tests and between occasions, and therefore also the interaction between occasions and tests. Hence there now remains only one main source of variation (the differences between the mean marks obtained by the several persons—the result in which we are chiefly interested) and two possible types of interaction. The effects of these three components are shown in Table VII.

TABLE VII. ANALYSIS OF VARIANCE FOR STANDARDIZED MARKS

Source of Variation	Sum of Squares (<i>Q</i>)	Degrees of Freedom	Mean Square (<i>V</i>)	Variance Ratio	Probability
P. Persons	6.0667	5	1.2133	22.39	<0.001
PT. Persons and Tests	.7038	20	.0352	—	—
PO. Persons and Occasions	2.1451	5	.4290	7.91	<0.001
PTO. Residual	1.0844	20	.0542	—	—
S. Total	10.0000	50	(.2000)		

As before, both the differences between the means for the persons and the interaction between persons and occasions are fully significant, the former exhibiting much the larger amount of variance. But the interaction between persons and tests has become so small that its mean square is even less than the mean square for the residuals.

We can now calculate the intra-class correlation as before from the square-sums for persons and for the interaction between pupils and trials. We obtain

$$r_{int} = \frac{Q_p - Q_{po}}{Q_p + Q_{po}} = \frac{6.0667 - 2.1451}{6.0667 + 2.1451} = 0.4776.$$

Since the degrees of freedom are now the same for both square-sums, the previous 'correction for bias' (substituting mean squares for square-sums) leaves the value for the coefficient unaltered.

The Factorial Analysis of Variance. We have seen that a reliability coefficient is intended to indicate the ratio of the estimated variance of the 'true' measurements to the actual variance of

¹ C. Burt, 'Factor Analysis by Submatrices', *Journ. Psychol.*, VI (1938), p. 349, Table I. To avoid a technical term which is apt to alarm students who are unfamiliar with matrix algebra, each 'submatrix' whose elements were to be summed was called a 'pooling square'. The phrase seems now to be more frequently used of the entire correlation matrix as thus partitioned.

Test Reliability Estimated by Analysis of Variance

should have been inferred from the attempt to determine the significance of the interaction between persons and tests in Table VII. But, since several correspondents have been puzzled by the occasional appearance of a negative result at this stage, it seemed useful to illustrate how it may arise.

TABLE VIII. ESTIMATES FOR THE VARIANCES OF THE POSTULATED FACTORS

Variance	Equation	Estimate
s_{γ}^2	$= \frac{1}{mn}(V_p - V_{pt} - V_{po} + V_{pto})$	·0803
s_{τ}^2	$= \frac{1}{m}(V_{pt} - V_{pto})$	-·0095
s_{ω}^2	$= \frac{1}{n}(V_{po} - V_{pto})$	·0750
s_e^2	$= V_{pto}$	·0542
$\sum s^2$	$= V_s$	·2000

Secondly, let us apply the foregoing analysis to the interpretation of the ordinary reliability coefficient. We now have¹

$$\begin{aligned}
 r_{xx'} &= \frac{V_p - V_{po}}{V_p + (m-1)V_{po}} = \frac{ns_{\gamma}^2 + s_{\tau}^2}{n(s_{\gamma}^2 + s_{\omega}^2) + s_{\tau}^2 + s_e^2} \quad \text{. (xiii)} \\
 &= \frac{.4051 - .0095}{.4015 + .3750 - .0095 + .0542} \\
 &= .4776 \text{ as before.}
 \end{aligned}$$

An equation like the above, expressed in terms of what may be called factorial variances, yields, I venture to think, a far clearer conception of what such coefficients are measuring than the more familiar equations which relate them to the cruder results furnished by an ordinary analysis of variance.² The detailed figures here obtained, for instance, bring into sharp relief the question raised above. When we speak of the 'reliability' of a test, we imply that it is a more or less 'reliable' measure of something; but of what? Are we seeking a reliable measure merely of the basic factor (γ) common to all the tests? Or are we intending to secure a more general estimate, which would include the specialized abilities of the separate tests as well?³ In other words, is τ (if significant) part of the 'true measurement', or is it irrelevant and therefore part of the 'error'? In the former case the figure obtained for the reliability would here be too low; in most cases it would be too high.

Reliability determined from a Single Application. There have been many attempts to derive an estimate of reliability, or at least of its upper or lower limits, from the results obtained by administering a composite test (or battery) once only. I have discussed this problem more fully elsewhere,⁴ and therefore need say very little here.

The relations between this procedure and the older and more familiar method may be most conveniently explained in terms of the pooling square. At first sight the idea that the consistency of two or more test-applications can be estimated from a single application only seems highly paradoxical.

¹ Since this section is intended to be of general application, I here give the general form of the equation for the unbiased intra-class correlation (see *Factors of the Mind*, p. 275, eqn. ii). Since in the present example $m = 2$, the equation reduces to the forms already used above (pp. 109 and 113).

² The relations between the two types of component are very similar to the relations between the summational or centroid factors obtained by 'simple summation' and the group factors obtained by a 'group factor analysis': with the former the 'general factor' includes not only the 'basic' factor but also contributions from the group factors, and similarly with the successive bipolars. This is confirmed by the results of an actual factor analysis by these two methods: see below, p. 118.

³ For example, when using a battery of verbal and non-verbal intelligence-tests, our aim is generally to measure intelligence only; when using question-papers in arithmetic, English, and the like to measure 'general educational ability' in a scholarship examination, our aim generally is to assess a compound of all the child's educational attainments. Elsewhere I have tried to distinguish the various types of so called 'reliability coefficient', and to give them more distinctive names: (see also Mr. Mahmoud's paper below, pp. 127f.).

⁴ See Burt, 'The Reliability of Teachers' Assessments', *Brit. J. Educ. Psychol.*, XV (1945), esp. Appendix, pp. 90-2.

It will be remembered, however, that our definition treats reliability as essentially a characteristic of the test, and accordingly, in estimating its value from two applications, we have assumed that the persons do not change significantly in the interval and that the test or subtests applied on the second occasion do not differ significantly from those applied on the first, or, if there *are* any changes in the persons or any differences in the tests, then they must be eliminated (so far as possible) by an appropriate statistical formula. Now, if there were no changes and no differences *whatever*, then the inter-correlations between the subtests on the second occasion would be identical with those obtained on the first, and the cross-correlations between two applications of the same test would take the form of 'reduced self correlations', i.e., their amount would be determined solely by the factors common to all the subtests, and would be unaffected by factors specific to either the first trial or the second.

With these assumptions we can now construct a suitable pooling square. The N.W. quadrant will, as before, contain the observed inter-correlations obtained from the first application, which is now the only *actual* application: as usual, it will have units (1.00) in the leading diagonal. The S.E. quadrant will represent the results of an *imaginary* second application carried out with perfectly equivalent subtests and under perfectly identical conditions. Hence all the correlations in the S.E. quadrant will be exactly the same as in the N.W. quadrant, and the side correlations in the N.E. quadrant will be exactly the same as the side correlations in both the N.W. and the S.E. quadrants: in the N.E. quadrant, however, the leading diagonal will no longer contain units, but 'reduced self correlations' conforming with the assumptions just laid down. The only question is—what values are we to substitute?

If we already possess estimates for the reliability of the separate subtests, we might perhaps use these; and their combined unreliability we may reasonably equate with s_e^2 . In this case the numerator for the reliability-ratio will differ from the denominator only by the omission of the term s_e^2 . We shall have in fact

$$r_{xx'} = \frac{\sum R_{ab}}{\sum R_{aa}} = \frac{n(s_\gamma^2 + s_\omega^2) + s_\tau^2}{n(s_\gamma^2 + s_\omega^2) + s_\tau^2 + s_e^2} \quad \text{. (xiv)}$$

Evidently this procedure will yield a higher value for the reliability than that obtained from two *actual* applications: for in the numerator of equation (xiii) (p. 116) the term ns_ω^2 was also omitted: when the requisite conditions are fulfilled, therefore, we may regard eqn. xiv as indicating an upper boundary.

However, as in factor analysis, it would seem more in accordance with general theory to eliminate from the diagonal elements not only the unreliability of each subtest but also its 'specific factor' (i.e., the group factor common to the two trials). Indeed, this is all we can hope to attempt if we possess no previous estimates for the unreliability, since in that case s_e^2 will be merged with s_τ^2 . We must therefore reduce the numerator of our equation to $n(s_\gamma^2 + s_\omega^2)$. To calculate this value, in the absence of any other empirical guide, the simplest practical device will be to take the average of the side-correlations as providing the best estimate for the average of the reduced self correlations. In terms of the hypothetical factor-variances, the formula for the reliability coefficient will then be

$$r_{xx'} = \frac{n(s_\gamma^2 + s_\omega^2)}{n(s_\gamma^2 + s_\omega^2) + s_\tau^2 + s_e^2} \quad \text{. (xv)}$$

Since, in the absence of all evidence to indicate their relative proportions, the distinction between s_γ^2 and s_ω^2 and between s_τ^2 and s_e^2 now becomes meaningless, we may re-write the formula in the simplified form¹

$$r_{xx'} = \frac{ns_G^2}{ns_G^2 + s_E^2} \quad \text{. (xvi)}$$

It is tempting to regard this as indicating the lower boundary. It must be remembered, however, that the factor G will now very probably incorporate most of the former factor ω and possibly much of the former factor τ . Hence, the estimated reliability may still be magnified to an unknown extent by the inclusion in the 'true' common-factor variance of contributions due to the special conditions obtaining at the time of the single trial. This obvious inference appears to have been ignored by many who advocate the simplified procedure.²

¹ This is the equation reached in the *Appendix* just cited (p. 90, eqn. v). The effect of using the average intercorrelation as described above is there indicated in eqn. viii; and the worked example given in the body of the article illustrates the relation of the coefficient so obtained to the results of the analysis of variance.

² Spearman's 'split-half' method is a special instance of this method, and is open to the foregoing (as well as other) objections. As I have shown elsewhere, the procedure described in the text and the formulae so deduced give the average value of the coefficients obtained by *all possible splits*.

Test Reliability Estimated by Analysis of Variance

The point is clearly illustrated by the present data, where the two trials were separated by an interval of 12 months (not, as our assumptions require, by the minimum time needed to prevent the effects of the first trial influencing the second). The sum of the 10 inter-correlations (without the self-correlations) obtained at the first trial was 17.09. Accordingly with the first method (inserting actual reliabilities for subtests) we obtain

$$\frac{17.09 + 1.77}{17.09 + 5 \times 1.00} = \frac{18.86}{22.09} = .854;$$

with the second (inserting estimates based on the average intercorrelation)

$$\frac{17.09 + 5 \times .85}{17.09 + 5 \times 1.00} = \frac{21.34}{22.09} = .966.$$

Reliability determined from Weighted Factors. The factor measurements obtained or implied by an analysis of variance consist of the simple sums or means of the observed test-measurements, without any differential weighting. As I have argued elsewhere, better estimates of reliability can be secured if we employ weighted sums, the appropriate weights being determined by factor analysis. Here, for example, the summational and the group factor analysis alike reveal that, on both occasions, Test *P* has a far higher loading for the general or basic factor than Test *T*; and this in turn greatly augments the size of the interaction between tests and persons.

It is largely for this reason that I have repeatedly urged the need, in any exact inquiry, to base the estimates of reliability on an explicit factorial analysis by a summational or group factor method.¹ But other questions frequently arise which only a factor analysis can solve. In an ordinary analysis of variance the most suitable form of classification is often a matter of some doubt—a point that is commonly overlooked. The ‘degrees of freedom’ can always be subdivided into ‘orthogonal comparisons’ in many different ways: and, as Mather observes, “the first task is to partition these degrees of freedom into the components that are appropriate to the various comparisons which might be interesting”.² But how can we tell which comparisons “might be interesting”? Shall we examine those that represent the special abilities conceivably required for particular groups of tests? and if so which groups do we select? Shall we look for those that represent ‘function-fluctuation’ in the persons, or, it may be, fluctuations in the conditions obtaining at successive trials, which themselves may affect different persons differently? Or again how are we to analyse the interactions into their constituent components and assess the importance of each? With large groups, or numerous subtests, or numerous repetitions, such problems often prove crucial. And nothing but a preliminary factor analysis can show how they are to be answered or which of the supplementary classifications are really worth while.

Since our initial table of data was based on so small a sample and does not represent any actual testing, I need not here reproduce the tables of correlations and factor saturations. Moreover, the method has been fully illustrated in previous papers.³ For completeness I will merely summarize the final results. Three summational factors and three group factors were extracted. Their contribution to the total variance (10.00) was as follows:

(A) General factor, 6.11; first bipolar (classifying occasions), 2.37; second bipolar (classifying tests), 0.35 (non-significant);

(B) Basic factor, 4.48; group factor for the first occasion, 1.88; group factor for the second occasion, 2.33.

It will be seen that the variances of the general factor and the bipolar factor for occasions stand in approximately the same proportions as the crude variances V_p and V_{po} in Table VII; while the basic factor and the two group-factors for occasions stand in almost exactly the same proportions as s_y^2 and s_w^2 . This is what, if the foregoing theory is correct, we should expect. So far as it goes, these points of agreement confirm the close relation between the principles underlying factor analysis and those underlying the analysis of variance. However, in many actual researches, where the diagnostic values and consequently the weights of the several subtests differ widely from one to another the agreement no longer obtains; and in such cases a reliability coefficient based on a factorial analysis will manifestly be far more accurate and far more informative than one derived by the more usual over-simplified procedures.

¹ Which factorial method is to be preferred will depend on the investigator's conception of the factor or factor measurements whose reliability he desires to assess. Thus I should not entirely agree with Mr. Mahmoud who appears to hold that the group factor method alone is appropriate for *all* such problems.

² *Op. cit.*, p. 73. Kendall observes that the choice is ‘as a rule determined by some prior classification given among the data of the problem’ (*Advanced Theory of Statistics*, II, 175). But in psychology, instead of being given *with* the data, the classification has often to be deduced *from* the data; and for this purpose a factorial analysis may be indispensable.

³ An excellent illustration is provided in this issue by the tables included in Mr. Mahmoud's article. If any reader wishes to use the correlational tables relating to the present data for purposes of class-instruction, copies can be obtained on request.

TEST RELIABILITY IN TERMS OF FACTOR THEORY

By A. F. MAHMOUD

The Need for a More Adequate Theory. In a previous issue of this *Journal*¹ I raised the question of assessing the reliability of composite tests (like the newly constructed battery of intelligence tests on which my colleagues and I have been engaged); and the Editor in reply drew attention to the possibility of using more refined techniques, such as have recently been developed from the familiar principles of factor analysis and analysis of variance. In several published researches one or other of these newer methods has been tentatively adopted. But the investigators have usually failed to extract all the information implicit in their data, either because the methods have not been carried far enough, or because they have been applied too mechanically and without an adequate recognition of the assumptions or hypotheses on which the procedures are based. It is the object of this note to restate the underlying theory² a little more simply, and then to show, by an actual example, first the value of the supplementary device that Professor Burt has termed the 'factorial analysis of variance', and secondly the superiority, for this particular problem, of a 'group factor analysis' rather than a summational or centroid procedure.

Illustrative Example. For the purposes of the inquiry referred to in my note, eight intelligence tests were constructed, and arranged to form two parallel batteries with similar subtests in each. Each of the subtests consisted of 20 problems or tasks of increasing difficulty. As they were intended for a comparatively illiterate population, all were of a non-verbal or performance type. Test 1 consisted of a picture-completion test—four stories each containing five panels in series with an inset to be chosen and inserted in each; Test 2 of geometrical designs to be constructed with coloured blocks, similar to Kohs'; Test 3 of 20 mazes, similar to the Porteus series; Test 4 of diagrammatic 'matrices' similar to Raven's series. A hundred boys were selected as forming a random sample of the population, and were tested with the first battery as nearly as possible on their eleventh birthday and again with the second battery some six months later. Thirteen were absent on the second occasion, so that the data analysed here relates to 87 individuals only. In scoring the test, arbitrary marks of 2 to 5 were awarded for each task correctly performed, the amount for each being determined partly by its complexity and partly by the testee's speed. Taking the unweighted sum of each pupil's marks for each subtest as his 'final mark' for the battery, and accepting the teachers' assessments for intelligence as a provisional criterion, we obtained validity coefficients of 0.56 for the first trial and 0.63 for the second—an average of 0.59. This is a somewhat disappointing figure. But as we shall see later on, better correlations are obtained when the subtests are given different weights; and, since the children were drawn from different schools and the teachers' standards varied appreciably, the criterion does not provide very trustworthy estimates of individual differences, and can only be regarded as furnishing a convenient overall check.

(a) *Analysis of Variance.* As a preliminary it seemed advisable to determine (i) whether the subtests comprising the two batteries showed any appreciable differences either in the general difficulty of the tasks or in the general method of marking, and (ii) whether the abilities that we sought to test showed any appreciable change with time.³ We therefore began by making a complete three-way

¹ 'A Reliability Coefficient based on Analysis of Variance', *Brit. J. Statist. Psychol.*, VII (1954), p. 61.

² I am much indebted to Professor Burt for supplying copies of his original memorandum on Reliability of Tests and Examinations, drawn up for the International Institute Examinations Enquiry, and of his later Laboratory Notes, and for permitting me to incorporate in my own presentation whatever extracts I wished. His abridged 'three-way analysis of variance' and what he calls 'the factorial analysis of variance' appear to be especially appropriate to problems like my own; and, save for a few unpublished theses and reports, do not seem to have been used in any large scale research.

In adapting his arguments, a few modifications have been made, chiefly in the interests of simplification: his notation, for example, distinguishes between (a) constants (or 'parameters') which characterize the population, whether actual or hypothetical, and (b) the corresponding estimates (or 'statistics') which are obtained from the sample. Here, to avoid complicating the discussion too much for the elementary reader, I have abandoned this double symbolism. The statistical reader will easily be able to make such distinctions for himself where he considers them important.

³ As will be obvious later on, it is scarcely possible to distinguish between the two explanations for the change in the averages, unless a more complicated experimental design is used. For such a purpose we should need at least four groups of children—one pair taking one battery first, the other pair taking the other battery first.

Test Reliability in Terms of Factor Theory

analysis of the variance for the raw scores. The working method followed was that recommended by Professor Burt in his Notes on *The Construction and Standardization of Tests* (illustrated in the preceding paper).

In analysing such data there are three main sources of variation to be considered—the persons tested (P), the tests employed (T), and the occasions on which the tests were administered (O). The crude analysis, based on the raw unstandardized scores, indicated (i) that the differences between the means for the several persons were highly significant; (ii) that the differences between the means for the several subtests were non-significant; (iii) that the differences between the means obtained on the two successive occasions were barely significant at the 5 per cent. level ($P = 0.047$): there was in fact a slight increase in the general average, due, it appeared, not so much to mental development or to previous practice as to slight differences in the standardization.

Accordingly, it seemed reasonable to eliminate these minor differences between tests and occasions as irrelevant, and reduce all scores to standard measure. In computing the standard deviations from the raw scores, the observed sums of squares were divided by ($N - 1$), not by N . The square-sum for each subtest will therefore be 86 and the mean zero. A fresh analysis of variance was then carried out. The results are shown in Table 1.

Since we have eliminated all differences between the means for subtests and for trials, and consequently the interaction between subtests and trials, we now have to deal with only four sources of variation, and the total number of degrees of freedom is reduced to

$$(N-1)mn = (87-1) \times 4 \times 2 = 688.$$

With the degrees of freedom indicated for the separate sources of variation, all three variance ratios appear to be fully significant even at the 0.1 per cent. level.

TABLE I. ANALYSIS OF VARIANCE

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
Persons (P)	480.069	86	5.5822	46.7
Persons and Subtests (PT)	129.748	258	.5029	4.2
Persons and Trials (PO)	47.326	86	.5503	4.6
Residual (PTO)	30.857	258	.1196	—
Total	688.000	688	(1.0000)	—

Note.—The value for each 'Mean Square' (in column 3) is obtained by dividing the 'Sum of Squares' by the 'Degrees of Freedom': hence the figure in the last line is not a 'Total' of those above.

Using the formula given by Burt,¹ we can calculate the correlation between the two batteries (i.e., the ordinary reliability coefficient) from the variances in the table, without having to compute the product-sums. We obtain

$$r_{tt'} = \frac{V_p - V_{po}}{V_p + V_{po}} \quad \dots \quad (i)$$

$$= \frac{5.5822 - .5503}{5.5822 + .5503} = .8205, \text{ where}$$

V_p and V_{po} denote the mean squares for persons and for the interaction between persons and trials.

The figure thus obtained is no doubt quite promising for a first attempt with a new set of tests; but (assuming we can accept it as it stands) it is too low for us to consider the battery sufficiently reliable for immediate use. Moreover, the analysis of variance reveals a feature which the reliability coefficient ignores—namely, the large interaction between persons and tests. This certainly calls for further scrutiny. Does it represent one additional and unsuspected factor or several? And does its magnitude fairly reflect that of the disturbing influences to which it points?

Psychologists who have used analysis of variance seem generally to suppose that the variances thus assigned to the several 'sources of variation' can always be identified, just as they are, with the variances of the hypothetical components or 'factors' that we are seeking to determine.² As a

¹ *Factors of the Mind*, p. 275, remembering that k (as there defined, our m) = 2. Cf. *Marks of Examiners*, p. 255. [See also article, p. 109].

² Burt's comparative tables for 'Analysis of Variance with Standardized Marks' and 'Analysis of Variance with Factors obtained by Weighted Summation' (this *Journal*, I, 12 and 25, Tables VI and XV) have perhaps encouraged this view. With the data there tabulated the interaction between persons and tests and between persons and occasions could be expressed in terms of a single row vector, analogous to a row of 'factor measurements' for a single factor in either case, and the variances were directly compared with the variances obtained from a factor analysis of the usual kind. But this was no doubt due to the need for a short and simplified exposition. In his fuller analysis of the Examinations Council data he stresses the need for what he calls the 'further analysis of variance' to obtain the purified 'factorial variances'.

matter of fact, however, when the analysis stops at the usual point, the 'square sums' calculated for the larger sources of variation always incorporate contributions from the smaller sources (i.e., from the interactions and residuals), much as the larger factors that are first extracted with a summational or centroid method incorporate contributions from the lesser and later group factors. Hence, to estimate the true magnitude of the hypothetical components the analysis must be carried one stage further. The importance of this for the assessment of reliability will be clearer if we turn for a moment to the more familiar approach—the examination of the relevant correlations.

(b) *Correlational Analysis.* Few investigators, in constructing a new battery, attempt a formal analysis of variance; but nearly all proceed to calculate the correlations between the various subtests. When the battery has been administered on two or more occasions, the coefficients may conveniently be arranged according to the scheme adopted by Burt in carrying out a 'group factor analysis'. The figures obtained in the present research are set out in Table II. The correlation matrix, it will be seen, has been partitioned into four submatrices or 'quadrants'. The inter-correlations for the first trial are entered in the N.W. quadrant, those for the second trial in the S.E. quadrant, and the cross-correlations in the N.E. and S.W. quadrants.

TABLE II. CORRELATIONS BETWEEN TESTS: (POOLING SQUARE)

Tests	1	2	3	4	Total (1st Trial)	1	2	3	4	Total (2nd Trial)
Test										
1. Picture	[1.000]	.897	.625	.713	3.235	.881	.768	.604	.637	2.890
2. Block	.897	[1.000]	.541	.584	3.022	.825	.837	.539	.563	2.764
3. Maze	.625	.541	[1.000]	.751	2.917	.445	.347	.703	.557	2.052
4. Matrix	.713	.584	.751	[1.000]	3.048	.559	.462	.667	.670	2.358
Total (1st Trial)	3.235	3.022	2.917	3.048	12.222	2.710	2.414	2.513	2.427	10.064
1. Picture	.881	.825	.445	.559	2.710	[1.000]	.774	.641	.735	3.150
2. Block	.768	.837	.347	.462	2.414	.774	[1.000]	.572	.704	3.050
3. Maze	.604	.539	.703	.667	2.513	.641	.572	[1.000]	.728	2.941
4. Matrix	.637	.563	.557	.670	2.427	.735	.704	.728	[1.000]	3.167
Total (2nd Trial)	2.890	2.764	2.052	2.358	10.064	3.150	3.050	2.941	3.167	12.308

The coefficients in each quadrant have been summed, in order to show how the sums of the correlations between the subtests may be used to obtain the correlation between the two entire tests. The ordinary proof of the formula for this purpose, as given by Pearson, Kelley and others, is lengthy and difficult for the ordinary student to follow.¹ It may, however, readily be simplified. The final marks which we desire to correlate are simply the unweighted² sums or averages of the marks obtained by each boy in the four subtests on the first and second occasions respectively. Accordingly, let w' (a row vector) denote the summation operation $[1, 1, \dots, 1]$, let M_a (an $n \times N$ matrix) denote the matrix of marks obtained with the first set of tests applied on the first day, and M_b that obtained with the second set applied on the second day. Then

$$r_{ss'} = \frac{w' M_a M_{ab} w}{\sqrt{(w' M_a M_a' w) (w' M_b M_b' w)}} = \frac{w' R_{ab} w}{\sqrt{(w' R_{aa} w) (w' R_{bb} w)}} \quad \text{. (ii)}$$

¹ Equation (ii) is merely the well-known formula for the correlation between sums (T. L. Kelley, *Statistical Method*, 1923, p. 197, eqn. 1.47). Burt gives the proof quoted in the text to show how the adoption of matrix notation simplifies the older algebraic demonstration: here, it will be seen, a derivation, which with Kelley occupies nearly two pages, is condensed into a couple of lines.

² When weights are used to maximize this correlation, the result is the 'canonical correlation' between the two batteries: cf. C. Burt, 'Factor Analysis and Canonical Correlations', this *Journal*, II (1948), 95-106: the formula for the unweighted correlation, derived by this method, is given on p. 97, eqn. iii.

Test Reliability in Terms of Factor Theory

or, to adopt the notation of the articles just cited,

$$r_{ss'} = \frac{\sum R_{ab}}{\sqrt{(\sum R_{aa} \sum R_{bb})}} \quad \dots \quad (ii\ a)$$

Thus, with the present example, the square sums for the two batteries are given by the sums of the intercorrelations shown in the N.W. and S.E. quadrants respectively of Table II; while the product sum is given by the sum of the cross correlations in the N.E. (or S.W.) quadrant.

Hence the correlation between the two batteries will be

$$r_{ss'} = \frac{10.064}{\sqrt{(12.222 \times 12.308)}} = .8205.$$

It will be noted that this is exactly the same value as was obtained from the analysis of variance by using equation (i). An algebraic proof of their equivalence will be given in a moment. Our more immediate object is to factorize this matrix of correlations in terms of the hypothetical components corresponding to those sources of variation which the preceding analysis of variance has shown to be significant.

Correlation as a Ratio of Variances. Fisher,¹ when first introducing his method of analysing variance, begins by explaining how a coefficient of correlation can be expressed as a ratio of two variances, or in other words, as "that fraction of the total variance ($A+B$) which is due to the cause

(A) which the observations have in common": i.e., in symbolic form, " $\rho = \frac{A}{A+B}$ ". He then points out that, whereas "the value of B (the 'error variance' so-called) may be estimated directly from the variance *within* groups", the value found for A , the variance *between* groups, includes a part contributed by B . The argument that follows may be regarded as a generalization of this analysis to the case in which we are concerned with three sources of variation (in addition to 'error' variance).

Hypothetical Model. In carrying out such an analysis, and particularly in applying it to the determination of reliability, the most urgent need, in my opinion, is to keep clearly in mind the underlying assumptions on which the procedure is based and to consider how far they hold good for the data under examination. Strangely enough, these obvious precautions seem rarely to be observed. Let us therefore start by attempting an explicit formulation of the assumptions involved.

For our present purpose, we may most appropriately begin with the simplest of Burt's 'mathematical models' (as he terms them). It is based on the provisional hypothesis that any observed test-measurement, obtained in the course of an investigation like the present, can be regarded as the weighted sum of four kinds² of hypothetical component—(i) a factor or factors³ common to all the tests on all the occasions; (ii) a group factor characteristic of similar specimens of the same test (with n kinds of test there may be n such components); (iii) a group factor representing those conditions that may be peculiar to the same occasion: (when the tests are applied on two or more occasions there may be two or more such components); (iv) a specific factor peculiar to each single test 'error of measurement': (with n tests and 2 occasions there will be $2n$ such components).

Let us then suppose that we are dealing with a battery of n tests, and that these are applied in identical or parallel form on two (or more) different days; and let m_{ijk} denote the observed mark or measurement⁴ obtained by the i th pupil in the j th test on the k th day ($i = 1, 2, \dots, N$; $j = 1, 2, \dots, n$; $k = 1, 2, \dots, m$). Let x denote the hypothetical factor measurements; and let g, d, t and e denote the relative weights (or factor saturations) of the four kinds of factor. Let the subscript i denote the individual to whom the measurements refer, and the subscript j or j' the test to which the saturations refer, the subscript k being added where it is necessary to distinguish days. Then, for any given day, we may write

$$m_{ij} = g_j x_{gi} + t_j x_{ti} + d_j x_{di} + e_j x_{ji} \quad \dots \quad (iii)$$

¹ *Statistical Methods for Research Workers*, chap. VII, sect. 40.

² This is merely a special case of the 'four factor theory' set out by Burt in his discussion of reliability in *Marks of Examiners*, p. 259, eqn. 5.

³ In the algebraic equations to be developed from these assumptions I shall for simplicity assume that there is only *one* component common to all the tests, or that, if there are more, their variances can be summed and treated as the variance of a single composite component. This simplification is warranted by Assumption 1: However, it means that, if there are factors common to two or more tests in the same battery but not to all, our present model must treat them as bipolar factors extending over all the tests (as is done in an ordinary centroid analysis). We could, of course, avoid this by introducing (as Burt does in one of his more elaborate models) a fifth kind of factor, namely, 'group factors for special abilities'. But I am anxious not to obscure the exposition of the main argument by raising numerous minor issues.

⁴ It seems a little inelegant, but not, I think, confusing, to use m (with subscripts) for 'marks' or 'measurements' and m (without subscripts) for number of occasions or trials. These symbols have been adopted in most of the papers, etc., to which I shall here refer. Hence it seemed best to retain them.

Following the usual convention, let us suppose that both the test-measurements and the hypothetical factor-measurements have been reduced to unitary standard measure, so that for any given day $\sum_i m_{ij}^2 = 1$, $\sum_i x_{gi}^2 = 1$, etc.

Assumption 1. In dealing with observable classifications (such as are used in analysis of variance) it may often happen that the crude *empirical* factors (i.e., the principles of classification) are not altogether independent of each other. For purposes of analysis, however, we shall assume that *the hypothetical components are strictly independent, i.e., mutually uncorrelated*.¹ Then all such product sums as $\sum_i x_{gi}x_{di}$, $\sum_i x_{ii}x_{ji}$, $\sum_i x_{ji}x'_{ji}$, etc., will vanish. Hence, for any given day,

$$\sigma_j^2 = \sum_i m_{ji}^2 = g_j^2 + t_j^2 + d_j^2 + e_j^2 \quad . \quad . \quad . \quad . \quad . \quad (\text{iv})$$

i.e., the variance of the j th test will be the sum of the variances of its four components. And similarly the correlation between two different tests applied on the same day will be

$$r_{jj'} = \sum_i m_{ij} m_{ij'} = g_j g_{j'} + d_j d_{j'}. \quad (v)$$

between two similar tests applied on different days will be

$$r_{j_1 j_2} = \sum_i m_{ij_1} m_{ij_2} = g_{j_1} g_{j_2} + t_{j_1} t_{j_2}. \quad (vi)$$

And between two different tests applied on two different days it will be simply

$$r_{j_1 j'_2} = \sum_i m_{ij_1} m_{ij'_2} = g_{j_1} g_{j'_2}. \quad \text{. (vii)}$$

The model we have adopted, together with this first assumption, will enable us to express the sums derived from the four quadrants of our 'pooling square' (cf. Table II above) in terms of four kinds of hypothetical component. Summing the n values for σ_j^2 and r_{jj} , and substituting the expressions given in eqns. (iv) to (vii), we obtain the results shown below in Table III (reproduced from Burt). It will be observed that, in virtue of the derivation of the factorial expansions for the various coefficients, the expressions $\sum t_{ji}^2$, $\sum e_{ji}^2$, $\sum t_{j_2}$, and $\sum e_{j_2}$ enter only into the diagonals of the N.W. and S.E. quadrants respectively, and the expression $\sum t_{j_1} t_{j_2}$ enters only into the diagonal of the N.E. and S.W. quadrants: in other words, these values are regarded as 'specifics' which merely augment the 'self correlations' in the matrices R_{aa} , R_{ab} , and R_{bb} . To determine the best values for each of the several factor saturations, a formal factor analysis by the group procedure will be essential (cf. Table VI below, p. 133). Meanwhile, if we introduce two further simplifications, we can reach a plausible approximation by means of a direct analysis of variance.

TABLE III. POOLING SQUARE WITH DIFFERENTIAL SATURATIONS
Sums of Correlations in terms of Hypothetical Saturations

Factor Saturations	Summed Correlations	
$g_{11} \ d_{11} \ - \ t_{11} \ - \ e_{11} \ - \ - \ -$ $g_{21} \ d_{21} \ - \ - \ t_{21} \ - \ e_{21} \ - \ - \ -$ $\dots \ \dots \ \dots \ \dots \ \dots \ \dots \ \dots \ \dots$	$\left. \begin{aligned} &\sum R_{aa} = \\ &(\sum_j g_{j1})^2 + (\sum_j d_{j1})^2 + \sum_j t_{j1}^2 + \sum_j e_{j1}^2 \end{aligned} \right\}$	$\left. \begin{aligned} &\sum R_{ab} = \\ &\sum_j g_{j1} g_{j2} + \sum_j t_{j1} t_{j2} \end{aligned} \right\}$
$g_{12} \ - \ d_{12} \ t_{12} \ - \ - \ - \ e_{12} \ -$ $g_{22} \ - \ d_{22} \ - \ t_{22} \ - \ - \ - \ e_{22}$ $\dots \ \dots \ \dots \ \dots \ \dots \ \dots \ \dots \ \dots$	$\left. \begin{aligned} &\sum R_{ba} = \\ &\sum_j g_{j1} g_{j2} + \sum_j t_{j1} t_{j2} \end{aligned} \right\}$	$\left. \begin{aligned} &\sum R_{bb} = \\ &(\sum_j g_{j2})^2 + (\sum_j d_{j2})^2 + \sum_j t_{j2}^2 + \sum_j e_{j2}^2 \end{aligned} \right\}$

¹ For a rigorous treatment it would be advisable to make it perfectly clear whether such conditions as this refer to the population or to the sample. That, in fact, is the procedure followed by Burt. But, as mentioned above, such a distinction entails a double set of symbols (Greek for the former, Roman for the latter) which is likely to bewilder the ordinary reader. It is perhaps sufficient to note that the *model* itself refers primarily to the population. Largely for convenience in subsequent deductions, certain modes of factor analysis also assume that the hypothetical factor measurements are uncorrelated in the sample. When the model is applied in any particular research, the investigator should first satisfy himself that the conditions postulated hold good within a reasonable approximation. For purposes of simplified calculation the requisite assumptions can then be supposed to be true also of the data contained in the sample.

Test Reliability in Terms of Factor Theory

Assumption 2. I shall next assume that *corresponding tests in the two batteries may be treated as 'equivalent'*. Burt's definition of 'equivalence' differs somewhat from that ordinarily put forward, and is perhaps a little too stringent. For any pair of tests to be regarded as 'equivalent' (he says), (i) their means and (ii) their standard deviations must be the same; (iii) they must depend on the same common factors (whether general and bipolar or basic and group); (iv) they must have the same weight for the same common factors and the same weight for the error factors (the error factors, of course, unlike the common factors, will not be the same).¹ From this it would follow that the corresponding self-correlations and intercorrelations in the N.W. and S.E. quadrants respectively differ only by such small amounts as may be attributable to errors of sampling. Evidently this is a corollary whose validity can be tested for any given set of data. In the present research, it will be seen, some of the observed intercorrelations do differ significantly; and these differences indicate obvious points at which the batteries require, if possible, to be improved. However, the totals for the two quadrants (12.222 and 12.308) differ hardly at all; and that may perhaps suffice to justify a provisional acceptance.

With this assumption the best estimate for the common standard deviation of the two batteries will be obtained by taking the arithmetic mean of the variances of both; and accordingly, as Burt suggests, we may now use the formula for the *intra-class correlation* (instead of the more familiar product-moment formula) together with the scheme for analysing variance associated with this coefficient.² In place of eqn. (ii a), therefore, we shall for the future substitute

r_{ss'} = \frac{\sum R_{ab}}{\frac{1}{2}(\sum R_{aa} + \sum R_{bb})} \dots \dots \dots (viii)

With the present data, the two sums, \sum R_{aa} and \sum R_{bb}, are so nearly equal that the substitution of the arithmetic mean for the geometric leaves the value for r_{ss'} (.8205 as given above) virtually unchanged.

Assumption 3. The summations carried out in an analysis of variance give equal weights to each component of the same type: in effect, the method assumes that g_{11} = g_{21} = \dots = \bar{g} (say), and d_{11} = d_{21} = \dots = \bar{d} (say), and so on, where \bar{g}, \bar{d}, etc., denote the average saturations. It is, of course, a familiar principle that, in summing a number of test-measurements, unless the differences between the weights are large, the substitution of equal weights has little effect on the relative values of the sums obtained (see this Journal, III, p. 111). Moreover, in constructing two parallel batteries of tests designed to measure the same ability, the investigator will, as a rule, discard any subtests which have a decidedly low weight for the general factor, or are likely to be more influenced than the rest by the conditions peculiar to a given trial. We may, therefore, now introduce this third assumption, which will still further simplify the resulting formulae and bring our factorization still more into line with the analysis of variance, namely, that, though the weights (or saturations) for different factors will usually be different, *the weights for the same factor are the same for all the tests*.

In addition then to what may be called the simple 'four factor hypothesis', our model will for the present assume (i) that the factors postulated are orthogonal, (ii) that as regards their factorial composition the test-batteries are equivalent, and (iii) that the saturations may differ from one factor to another but are the same for the same factor. In our present data there is nothing that flagrantly conflicts with these assumptions: but for a strict test of their accuracy a full factor analysis would be essential. This we shall attempt at a later stage.

We can now reconstruct the correlations to be expected on the basis of these last two assumptions. As will be seen from Table IV, the hypothesis on which our model is based treats the N.W. and S.E.

TABLE IV. POOLING SQUARE WITH EQUALIZED SATURATIONS
Sums of Correlations in Terms of Hypothetical Saturations

Factor Saturations						Summed Correlations	
$\bar{g}$	$\bar{d}$	$-\bar{t}$	$-\bar{e}$	$-\dots$	$-\dots$	$\left. \begin{aligned} \sum R_{aa} \\ = (n\bar{g})^2 + (n\bar{d})^2 + n\bar{t}^2 + n\bar{e}^2 \end{aligned} \right\}$	$\sum R_{ab}$ $= (n\bar{g})^2 + n\bar{t}^2$
$\bar{g}$	$\bar{d}$	$-\dots$	$-\bar{t}$	$-\bar{e}$	$-\dots$		
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$		
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$		
$\bar{g}$	$-\bar{d}$	$\bar{t}$	$-\dots$	$-\bar{e}$	$-\dots$	$\left. \begin{aligned} \sum R_{ba} \\ = (n\bar{g})^2 + n\bar{t}^2 \end{aligned} \right\}$	$\sum R_{bb}$ $= (n\bar{g})^2 + (n\bar{d})^2 + n\bar{t}^2 + n\bar{e}^2$
$\bar{g}$	$-\bar{d}$	$-\bar{t}$	$-\dots$	$-\bar{e}$	$-\dots$		
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$		
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$		

¹ Once again I have ventured to simplify. Burt distinguishes between 'equivalent tests' and 'parallel tests'. But this is unnecessary for our present purpose.
² See Fisher, *loc. cit. sup.*, pp. 199 f., and Tables 38 and 39.

quadrants as identical not only in their sums but also in their composition. Accordingly, since we now have $R_{bb} = R_{aa}$, we can write

$$r_{ss'} = \frac{\sum R_{ab}}{\sum R_{aa}} = \frac{(\bar{ng})^2 + n\bar{t}^2}{(\bar{ng})^2 + (\bar{nd})^2 + n\bar{t}^2 + n\bar{e}^2} = \frac{A}{A+B}, \quad \text{. . . . (ix)}$$

in Fisher's notation. Thus interpreted, the ordinary reliability measures *the ratio of (a) the variance resulting from the general factor and the factors specific to similar tests to (b) the total variance, which includes in addition the variance due to the different occasions and to the errors of measurement.*

Is this what we really want to measure when we are estimating the reliability of our battery?

Computation of Variances. Before we take up this question, it will be well to consider whether it is really possible to estimate the values of these averaged factor saturations, and if so how. As a matter of fact two alternative procedures are available. With standardized measurements the *average inter-correlation* for the separate tests is identical with the *intra-class correlation*, and an unbiased estimate of this can readily be derived from the variances.¹ Hence we can start either from the variances or from the detailed correlations.

(a) *The Factorial Analysis of Variance.* From the symbolic tables set out above, it is obvious that what is wanted are not the factor saturations as such, but merely their squares, i.e., the factor variances. These can readily be computed by the device which Burt has called the 'factorial analysis of variance'—a further analysis in which the crude variances, computed for an ordinary analysis of variance, are themselves re-analysed to obtain, as it were, the purified variances for the several hypothetical factors that express the classification employed.

The requisite formulae are given in the earlier memorandum already cited. Using V to denote a crude variance derived by an ordinary analysis of variance (as in Table 1 above) and σ^2 to denote the variances of the hypothetical factors, we have the following equations, and then, inserting the values for V given in Table 1, we obtain the figures shown on the right:

$$\begin{aligned} \sigma_p^2 &= (V_p - V_{pt} - V_{po} + V_{pto}) \div mn && \cdot 5811 \\ \sigma_{pt}^2 &= (V_{pt} - V_{pto}) \div m && \cdot 1916 \\ \sigma_{po}^2 &= (V_{po} - V_{pto}) \div n && \cdot 1077 \\ \text{and } \sigma_{pto}^2 &= V_{pto} && \cdot 1196 \end{aligned}$$

$$\sigma_p^2 + \sigma_{pt}^2 + \sigma_{po}^2 + \sigma_{pto}^2 = V_s \quad \quad \quad 1 \cdot 0000$$

It will be remembered that the marks for each subtest were reduced to standard measure. The square-sum for a standardized subtest is $N-1$ (the number of degrees of freedom in the sample of persons), i.e., with the present sample 86. Since there are $m \times n = 8$ subtests, the total square-sum must be $86 \times 8 = 688$. But this was also the figure obtained for Q_t by adding the sums of squares in column 1 of Table I. Re-dividing that total by the number of degrees of freedom $(N-1)mn$, we obtain for the average 'mean square' of any one measurement the value of 1.000 (the figure for V_s in the table above). These results provide a useful check on the computer's arithmetic.

The four values given in the table above show how this average variance may itself be subdivided to show the proportionate contributions from the four hypothetical factors. The basic factor contributes over 58 per cent.—a decidedly encouraging result; the supplementary factor for tests contributes over 19 per cent.—an unduly large amount; and the two smaller factors, representing susceptibility to the influences of the occasion and error of measurement in the narrower sense, together contribute another 22 per cent. or rather more. Evidently, before attempting to revise the test material, it will be desirable to investigate more closely the precise causes of these undesirable fluctuations: to that we shall turn in a moment.

(b) *Correlational Factors.* From the formulae already summarized in Table IV we can deduce the relations between the factors obtained by this 'further analysis of variance' and those obtained by the 'sum method' of factorizing the detailed correlations.² The method of deduction will be

¹ Cf. C. Burt, *Factors of the Mind*, 1940, p. 275, and the I.I.E.C. memoranda referred to above.

² Cf. this *Journal*, II (1949), 62 (appendix).

Test Reliability in Terms of Factor Theory

simplified if, as Table IV suggests, we first average the intercorrelations for the several subtests. We can thus re-write the values in the pooling square as follows :

			Total			Total		
1	$\bar{r}_s$	...	$1+(n-1)\bar{r}_s$	$\bar{r}_{sd}$	$\bar{r}_d$	...	$\bar{r}_{sd}+(n-1)\bar{r}_d$	
$\bar{r}_s$	1	...	$1+(n-1)\bar{r}_s$	$\bar{r}_d$	r_{sd}	...	$\bar{r}_{sd}+(n-1)\bar{r}_d$	
...	...	...	...	...	...	...	...	
Total			$n[1+(n-1)\bar{r}_s]$				$n[\bar{r}_{sd}+(n-1)\bar{r}_d]$	
$\bar{r}_{sd}$	$\bar{r}_d$	...	$\bar{r}_{sd}+(n-1)\bar{r}_d$	1	$\bar{r}_s$	...	$1+(n-1)\bar{r}_s$	
$\bar{r}_d$	$\bar{r}_{sd}$	...	$\bar{r}_{sd}+(n-1)\bar{r}_d$	$\bar{r}_s$	1	...	$1+(n-1)\bar{r}_s$	
...	...	...	...	...	...	...	etc.	

Or, substituting the numerical averages from Table I,

			Total				Total
1.0000	.6888	...	3.0664	.7727	.5811	...	2.5160
.6888	1.0000	...	3.0664	.5811	.7727	...	2.5160
...	...	...	...	...	...	...	...
—	—	—	12.2656	—	—	—	10.0640
...	...	...	...	...	...	...	...

If (as here) the tests have been standardised so that $\sum \sigma^2 = 1.000$, then the values for the factor variances will be given by the following equations :

$$\begin{aligned}
 \bar{g}^2 &= \sigma_p^2 = \bar{r}_d & &= .5811, \\
 \bar{t}^2 &= \sigma_{pt}^2 = \bar{r}_{sd} - \bar{r}_d & &= .7727 - .5811 = .1916, \\
 \bar{d}^2 &= \sigma_{po}^2 = \bar{r}_s - \bar{r}_d & &= .6888 - .5811 = .1077, \\
 \bar{e}^2 &= \sigma_{pto}^2 = 1 - \bar{r}_{sd} - \bar{r}_s + \bar{r}_d = 1 - .7727 - .6888 + .5811 = .1196.
 \end{aligned}$$

It will be seen that these equations yield exactly the same values as before. They are in fact merely a simplification of the ordinary method of group factor analysis, adapted to fit the hypothetical scheme of basic, group, and specific factors that we postulated at the outset.

Various Forms of the Reliability Coefficient. Now, as has frequently been pointed out, there are in practice several alternative ways of estimating the so-called reliability of a test ; and these different procedures are by no means equivalent. Indeed, they imply quite different definitions for what is meant by reliability ; and much confusion would be avoided if they were given different names. Three main forms¹ are commonly distinguished ; and, as Burt has observed, they may most conveniently be expressed in terms of factors such as we have just described. Slightly modifying his definitions to bring them into line with those of later writers, we may formulate them as follows.²

¹ E.g., J. P. Guilford, *Psychometric Methods* (1936), p. 411 ; H. Gulliksen, *The Theory of Mental Tests* (1950), pp. 193 f.

² One of the earliest publications which uses factor-theory in a discussion of reliability is the article by Miss B. M. D. Cast on marking English composition (*Brit. J. Educ. Psychol.*, IX, 1939, pp. 257 f., X, 1940, pp. 49 f.). Her procedure has been largely followed in Ebbelwhite-Smith, *Marking English Essays* (1941), pp. 42 f. ; and was itself based on Burt's earlier L.C.C. Memorandum. Pilliner has similarly adopted Burt's 'three-way analysis of variance' to describe the correlations between marks awarded by different markers to different essays written by the same pupils ; but apparently he does not accept the use of the intra-class correlation, and he retains the differences between the means for essays and the means for marks (A. E. G. Pilliner, 'Applications of Analysis of Variance to Problems of Correlation', this *Journal*, V, 1952), pp. 31-8). H. W. Alexander, like Burt, makes full use of the intra-class correlation ('The Estimation of Reliability when Several Trials are available', *Psychometrika*, XII, 1947, pp. 79-100). His variance equations for the biased and unbiased forms are the same as Burt's (cf. *Factors of the Mind* (1940), pp. 274 f.) and so is his formula showing the relation between the intraclass correlation and the average correlation.

An instructive discussion of the problems raised, which is also expressed in terms of factor analysis (viz., Holzinger's bifactor theory) will be found in L. J. Cronbach, 'Test Reliability : Its Meaning and Determination' (*Psychometrika*, XII, 1947, pp. 1-16). The algebraic expressions there proposed are not explicitly related to the techniques of the pooling square or the analysis of variance : and the factors refer

Test Reliability in Terms of Factor Theory

must allow for the presence of both additional sources of variation. But are we also to include them in the numerator of the ratio, or should we endeavour to eliminate them? The answer depends on what precisely we desire to measure. At least three different replies are possible. Hence we have to consider three further types of reliability coefficient and three further formulae.

Definition 2. The 'Coefficient of External Consistency' or 'Equivalence' measures the degree to which a composite test, applied on a second occasion, either in the same or in a closely parallel form, yields results agreeing with those of the first application, even though the conditions of testing may differ. If the tests are identical, we may speak of 'consistency'; if not, of 'equivalence'.¹ For this purpose we evidently require to introduce $\bar{t}^2$ into the numerator and both $\bar{t}^2$ and $(n\bar{d})^2$ into the denominator. The formula will therefore be that already reached (eqn. ix) when calculating a test-retest correlation from the pooling square, namely:

$$r_e = \frac{n\bar{g}^2 + \bar{t}^2}{n\bar{g}^2 + n\bar{d}^2 + \bar{t}^2 + \bar{e}^2} = \frac{n\sigma_p^2 + \sigma_{pt}^2}{n\sigma_p^2 + n\sigma_{po}^2 + \sigma_{pt}^2 + \sigma_e^2} \quad \text{. (xiii)}$$

This is the form to which eqn. ix reduces when n is cancelled from both numerator and denominator. We may therefore regard it as the variance equivalent of the familiar 'test-retest correlation' with parallel forms.

Substituting the values already obtained, we have

$$r_e = \frac{4 \times .5811 + .1916}{4 \times .5811 + 4 \times .1077 + .1916 + .1196} = \frac{2.5160}{3.0664} = .8205.$$

Or, using the average correlations $\bar{r}$ as they stand,

$$r_e = \frac{\bar{r}_{sd} + (n-1)\bar{r}_d}{1 + (n-1)\bar{r}_s} = \frac{2.5160}{3.0664} = .8205.$$

Nevertheless, instructive as this coefficient may be for preliminary inquiries, it does not appear to furnish precisely the information we require. Our tests have been designed to measure the common ability, g (whose variance is designated by σ_p^2): the more specific abilities peculiar to the several subtests (whose variance is designated by σ_{pt}^2) do not, as a rule, form part of the particular function we set out to test.³ We must therefore seek some way of eliminating these disturbing factors from the numerator. If we keep to a correlational technique, the appropriate device seems plain. Much the same difficulty confronted the earlier factorists; and to eliminate the irrelevant specifics from their correlation matrices, they proceeded to substitute 'reduced self-correlations' or reduced self-covariances for the self-correlations or observed test-variances entered in the leading diagonal. Here we can apply the same procedure either (i) to the submatrix of cross-correlations (R_{ab}), thus eliminating $\bar{t}^2$, or (ii) to the submatrix of intercorrelations (R_{aa}), thus eliminating $(\bar{t}^2 + \bar{e}^2)$. In ordinary factor analysis the values to be inserted would be reached by successive approximation; but if (as we are here assuming) the saturations for the same factor do not differ greatly in magnitude, we may plausibly treat the average value of the self-correlations as approximately equal to the average of the inter-correlations. To indicate that the submatrices now have reduced values in their diagonals, I shall follow the usual convention, and affix an asterisk.

By referring to Tables I and IV, the reader can readily verify for himself the fact that the correlational procedure, with these reductions, yields the same values as the variance procedure, with the proposed omissions. We have in fact

$$\begin{aligned} \sum R_{aa} &= n(n\sigma_p^2 + n\sigma_{po}^2 + \sigma_{pt}^2 + \sigma_e^2) = 4(2.3244 + .4308 + .1916 + .1196) = 12.2656; \\ \sum R_{aa}^* &= n(n\sigma_p^2 + n\sigma_{po}^2) = 4(2.3244 + .4308) = 11.0208; \\ \sum R_{ab} &= n(n\sigma_p^2 + \sigma_{pt}^2) = 4(2.3244 + .1916) = 10.064; \\ \sum R_{ab}^* &= n(n\sigma_p^2) = 4(2.3244) = 9.2976. \end{aligned}$$

¹ The term 'coefficient of consistency' was proposed by P. Hartog (*Marks of Examiners* (1936), p. xiv, footnote 1). The phrase 'coefficient of internal consistency' was suggested to denote more specifically the coefficient derived from data obtained at a single administration, i.e., from the intercorrelations of the subtests, and the phrase 'coefficient of external consistency' to denote the coefficient derived from two administrations.

² Cf. Burt, *Marks of Examiners*, p. 303, eqn. xl.

³ In some inquiries, particularly in the field of educational and vocational selection, the examiner may desire to test both general capacity and special aptitudes or knowledge. Thus, in the old junior county scholarship examination, tests involving reading, spelling, composition, and arithmetic were set with the intention of assessing not only the candidates' general educational ability (g) but also their special scholastic abilities and attainments (t). In such cases eqn. xiii remains the appropriate formula.

As a result of these various modifications, we obtain three further coefficients.

Definition 3. The 'Coefficient of Internal Consistency' measures the degree to which the results of the various *constituent* tests agree with each other, and therefore the degree to which the results of a second application, with a battery assumed to be internally identical with the first and under conditions assumed to be exactly the same, would agree with the results of the first. On referring to the symbolic correlation matrix (Table IV) it is evident that the theoretical formula required by the above assumptions will be

$$r_c = \frac{n(\bar{g}^2 + \bar{d}^2)}{n(\bar{g}^2 + \bar{d}^2) + \bar{t}^2 + \bar{e}^2} = \frac{\sum R_{aa}^*}{\sum R_{aa}} = \frac{11.0208}{12.2656} = .8985. \quad \text{. (xiv)}$$

The coefficient refers primarily to the internal consistency of the test as assessed by the results of a single trial.¹ But naturally, the figure will vary somewhat with the results of different trials. Here we have in effect averaged the consistency coefficient for two. If there is any likelihood that practice or familiarity may alter the internal consistency, then we must base our coefficient solely on the first trial. For that, the same procedure here yields $r_c = .8970$.

The equation can thus be employed when the tests have only been administered once; this of course is the attraction of all such formulae. However, with a single trial it is obviously impossible to distinguish between $\bar{g}^2$ and $\bar{d}^2$ or between $\bar{t}^2$ and $\bar{e}^2$: the two former are merged into a single factor common to all the four tests and the two latter represent the specific factors peculiar to the results obtained from a single test on a single occasion. Consequently the pooled variances may be rewritten

$$\frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2}; \quad \text{. (xv)}$$

and this expression, as Burt has shown, is equivalent to the 'split half' reliability coefficient obtained from a single trial *when all possible splits are taken and averaged* (this *Journal*, VI, 61).

Definition 4. The 'Coefficient of Test Stability' indicates the degree to which the measurements of an ability assessed by one set of tests remain the same (or would remain the same) when measured by a second set, even though the methods or conditions of testing may differ. The theoretical formula will be

$$r_{s/t} = \frac{n\bar{g}^2}{n(\bar{g}^2 + \bar{d}^2) + \bar{t}^2 + \bar{e}^2} = \frac{\sum R_{ab}^*}{\sum R_{aa}} = \frac{9.2976}{12.2656} = .7580. \quad \text{. (xvi)}$$

Assuming that g (the general factor as assessed by the unweighted means for the several persons) represents the characteristic that the battery of tests was designed to measure, this equation comes nearest to expressing what most psychologists appear to have meant by reliability. It will be seen that the formula can only be computed if data from at least two trials are available. The tests should be so compiled, and the administration so carried out, that both $\bar{d}^2$ and $\bar{t}^2$ are reduced to a minimum; and, when that has been done, the coefficient so calculated will probably approximate to the coefficient of internal consistency. But, unless the structure of the test and the peculiarities of the function tested (e.g., its liability to change) are already well known, it would be rash to state whether that coefficient indicates a lower or an upper bound.

Definition 5. The 'Coefficient of Person Stability' measures the extent to which the relative abilities of a given set of persons, assessed on two or more separate days, have remained the same, in spite of the interval between the two applications or (particularly if the interval is short) in spite of the variations in the conditions that obtained. For this the most obvious formula will be

$$r_{s/p} = \frac{\bar{g}^2}{\bar{g}^2 + \bar{d}^2} = \frac{\sum R_{ab}^*}{\sum R_{aa}} = \frac{9.2976}{11.0208} = .8437. \quad \text{. (xvii)}$$

This concept is not unlike that designated 'function stability' by certain writers.² But the explanations generally offered overlook an important qualification. A given function or ability may tend to change in *all* persons at approximately the same absolute or proportional rate: if, for example,

¹ "The determination of reliability from a single test-application" is fully discussed by Burt in 'The Reliability of Teachers' Assessments', *Brit. J. Educ. Psychol.*, XV, 1945, pp. 80-92, and the use of analysis variance illustrated in detail.

² Cf. G. B. Paulsen, 'A Coefficient of Trait Variability', *Psychol. Bull.*, XXVIII, 1931, pp. 218 f., and R. H. Thouless, 'Test Unreliability and Function Fluctuation', *Brit. J. Psychol.*, XXVI, 1936, pp. 325 f. I should add that the more important formulae given above are virtually identical with those suggested by Burt in his memorandum for the International Institute Examinations Inquiry: cf. *Marks of Examiners* (1936), p. 274, eqns. xii and xv. They are here given with a slightly different subscript-notation adopted in his *Laboratory Notes*.

Test Reliability in Terms of Factor Theory

the I.Q. were absolutely constant for every child, the mental ages of the children tested would alter, and alter by different annual increments, but the spaced order of merit would remain the same from year to year; the same would happen if all pupils increased in the knowledge of a given school subject in accordance with some constant educational quotient. This kind of change is not included in the foregoing concept. In actual fact, however, as numerous investigations have shown, the correlations between the I.Q.'s of the same group of pupils tend to diminish with increase in time; and this implies that the changes in mental age take place more or less erratically, so that the order of ability does not remain the same. Hence what we require to measure is the *relative stability of the persons* rather than the stability of the function or capacity in and for itself.

The Coefficient of Trait Variability. By subtracting any of the foregoing 'coefficients of reliability' from unity, we could if we wished obtain a 'coefficient of unreliability' (or 'error'). There is, however, one additional coefficient of unreliability that deserves special mention, namely, the 'Coefficient of Trait Variability'. This measures the degree to which the trait or ability fluctuates from one occasion to another. The amount of variation is indicated by subtracting the 'Coefficient of Test Stability' (i.e., stability of the test measurements from one day to another), not from unity, but from the 'Coefficient of Internal Consistency' (i.e., consistency of tests applied on the same day). This yields

$$r_{(tv)} = \frac{n\sigma_{po}^2}{n(\sigma_p^2 + \sigma_{po}^2) + \sigma_{pt}^2 + \sigma_{pto}^2} = .8985 - .7580 = .1405. \quad (xviii)$$

It is instructive to compare this with Thouless's formula for 'function fluctuation' (*loc. cit.*, p. 332),

$$I(f) = \frac{r_{(a_1 - a_2)(b_1 - b_2)}}{\frac{1}{2}(r_{a_1 b_1} + r_{a_2 b_2})} \quad (xix)$$

where *a* and *b* denote the two tests, and 1 and 2 the occasions on which they are applied. It is not easy to express this in terms of the constants we have so far used, because with only two tests and two occasions the variances of the tests and still more of their differences are liable to be highly erratic. However, making the same assumptions as before,¹ it becomes very approximately equivalent to

$$\frac{\sigma_{po}^2}{(\sigma_{po}^2 + \sigma_{pto}^2)(\sigma_p^2 + \sigma_{po}^2)} \quad (xx)$$

¹ Burt points out that Yule (in his discussion of attenuation due to unreliability) first suggested calculating the correlation $r_{(a_1 - a_2)(b_1 - b_2)}$ as a 'partial test' for the presence or absence of 'fluctuation', as later writers have called it (U. Yule, *Introduction to the Theory of Statistics* (1910), p. 214; Yule's method was used by Brown to compute the 'variability of function within the individual' from his own data, *Essentials of Mental Measurement* (1911), p. 84). But, as Burt observes, 'the precise grounds for the choice of the denominator (for eqn. xix) are by no means clear'. With the usual assumptions regarding equality of variances and independence of factors in the population (assumptions which may be quite flagrantly vitiated in small samples and with factors derived from unweighted means) he goes on to show that Thouless's coefficient is approximately equivalent to

$$\frac{(V_{po} - V_{pto})(V_p + V_{pt} + V_{po} + V_{pto})}{(V_{po} + V_{pto})(V_p - V_{pt} + V_{po} - V_{pto})} \quad (xx a)$$

which reduces to equation xx when (as here) the measurements are so standardized that

$$\sigma_p^2 + \sigma_{pt}^2 + \sigma_{po}^2 + \sigma_{pto}^2 = 1.00.$$

In his worked example he uses the constants obtained in his article on 'Factor Analysis and Analysis of Variance' (this *Journal*, I, p. 12) and obtains (a) with Thouless's formula $.031/.608 = .051$ and (b) with eqn. xx a $.038/.617 = .062$. It will be observed that eqn. xx, or its equivalent eqn. xx a, yields a value of zero when σ_{po}^2 is zero and a value of unity when V_p and V_{pt} are both zero. This version of the ratio would therefore seem better than the correlational form, but even so is not quite consistent. However, 'for any genuine assessment of the amount of instability or trend in the ability or "function" tested, more than two trials are essential; and the foregoing formulae are accordingly generalized to meet cases that involve more than two trials and more than two subtests in each trial. Where the trends are linear (whether produced by positive or negative practice effects due to the tests themselves or by mental growth or training), the ordinary methods of analysis of variance can be employed; if they are curvilinear or irregular, the line for trend can be fitted by the methods discussed by Philpott and Burt in their investigations of work-curves. But, for a really adequate investigation of the question, factorial methods, such as are used in the later part of the article, seem indispensable.

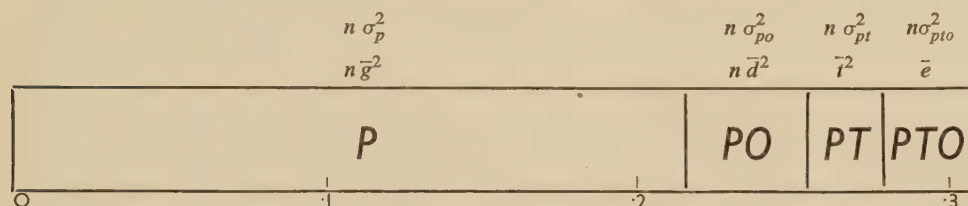
While the present paper is passing through the press, an article by C. C. Anderson on 'Simple Methods of Testing for Function Fluctuation' has appeared in *Brit. J. Psychol.*, XLVI, 1955, pp. 1-19. His 'model' is similar to Burt's, and from it he derives a formula for 'Thouless's criterion' in terms of 'mean squares'.

With the foregoing figures eqn. xx yields a value of 0.68; eqn. xix (after similar subtests have been pooled) yields one of 0.61. Both values seem far too high.

Comparative Values of the Various Coefficients. It is often supposed that the different methods of assessing reliability may be arranged in order according to the size of the coefficients they yield. For instance, several writers have stated that estimates based on single trials—estimates of 'internal consistency'—are higher than those based on two trials—the 'test-retest' procedure: the former it is said, avoids the attenuation due to different days and different conditions, but may be unduly enlarged when irrelevant influences operative on the same day affect *all* the subtests; the latter is diminished by the varying changes the examinees may undergo in the interval between the trials; and, if 'equivalent forms' have been employed, will be still further reduced by the lack of complete equivalence. Other writers have drawn the opposite conclusion. Jackson, for example, compiled five tests of intelligence, and found in every case that estimates based on test and retest were higher than those based on internal consistency (*Psychometrika*, VII, 159): such a result is to be expected when the compiler selects a heterogeneous set of problems in order to assess intelligence from different angles and perhaps deliberately discards those that are likely to duplicate each other.

A comparison of the foregoing formulae will show that inconsistencies like these are almost inevitable. The differences between one coefficient and another do not result from just adding yet another element in turn to the 'true variance': sometimes one element is substituted for another in the

FIG. 1.—Contributions of Component Variances to Total Variance.



numerator, and others are omitted or inserted in the denominator. In Figure 1 the magnitude of the four component variances as obtained in the present experiment is illustrated by a 'bar diagram'.¹

Unfortunately he gives no proof, and there are at least two misprints in the formula. Instead of what is there printed, we should probably read (in Anderson's notation) $\frac{B-C}{B+(p-1)C}$. But since in all his experiments $p = 2$, this simplifies to $\frac{B-C}{B+C}$. Anderson's formula is thus equivalent to omitting the division by .617 in eqn. (b) above, and seems to assume that the denominator in Thouless's ratio may be taken as approaching unity when the factors are genuinely independent, and as counteracting the effect of their correlations when they are not.

An alternative formula suggested by Burt is

$$I_{(iv)} = \frac{\sum R_{aa} - \sum R_{ab}}{n^2 + \sum R_{ab}} = \frac{n\sigma_{po}^2 + \sigma_{pto}^2}{n(\sigma_{po}^2 + \sigma_{pto}^2) + (n-1)\sigma_{pt}^2}$$

If the variance due to *occasions* is zero, this reduces to $\sigma_{pto}^2/[n\sigma_{pto}^2 + (n-1)\sigma_{pt}^2]$, which diminishes indefinitely as n increases indefinitely, and in any case becomes exactly zero if the error variance is also zero. If the specific variance due to the *tests* is zero, it reduces to $(n\sigma_{po}^2 + \sigma_{pto}^2)/(n\sigma_{po}^2 + n\sigma_{pto}^2)$, and becomes exactly unity if the error variance is also zero. With the present data the value of this 'index' would be .3710—a far more plausible figure.

The chief differences between $r_{(iv)}$ and $I_{(iv)}$ arise from the addition of the error variance in the numerator and the disappearance of the general factor variance from the denominator. This seems plausible, since the absolute size of the general factor variance is quite irrelevant to the question raised which the 'index' seeks to answer.

¹ The diagram is intended to represent a simple and instructive piece of apparatus used for demonstrations by Burt. A long and narrow tray (similar in form to the diagram) holds a sequence of four oblong panels, painted white on the front and black on the back. When the constituents are included in the numerator (the 'true variance'), the white side is left uppermost; when they are included only in the denominator (the 'total variance') and thus represent part of the 'error variance', they are turned over and the black side is uppermost; when they are excluded from both, they are removed altogether. Various sets of four such panels represent the results obtained from different types of experiment.

Test Reliability in Terms of Factor Theory

From this it will at once be seen that with the present data the substitution of σ_{pt}^2 for $n\sigma_{po}^2$ will reduce the value; with other data it may have the opposite effect; and similarly with the changes in the denominator. In short, the equations we have reached make it obvious that no fixed order of magnitude, valid for all kinds of data, can possibly be laid down.

Factor Analysis. The formulae which I have so far described may suffice the purposes of a preliminary inquiry. It must, however, be remembered that they were based on certain simplifying assumptions, introduced partly to permit the adoption of well recognized routine techniques, such as an ordinary analysis of variance, and partly to keep the working methods as short and easy as possible. Let us now examine the effect of removing some of the most questionable assumptions, and inquire how far the more elaborate modes of computation will alter the main results.

Throughout the foregoing calculations we have assumed that the weights (i.e., the 'loadings' or 'saturations') for any one factor may be treated as equal for all the tests. This assumption, as we have seen, is implicit in the formulae for the analysis of variance. Yet in practice it is seldom likely to be strictly true; and it is certainly conceivable that the appropriate weights might differ so widely as to render the inferences hitherto drawn decidedly precarious. The matrix representing the interaction between tests and persons itself, as was noted at the time, also deserves a more detailed treatment. Our simplified procedure assumed that the factors producing this interaction were merely isolated components specific to each test—in other words, that, apart from the figures in the diagonal, the correlational submatrices have unit rank.¹ But once again such an assumption may be quite unwarranted. If so, it becomes conceivable that our previous estimate for the reliability of the battery as a test of intelligence will turn out to have been unduly magnified by the inclusion of more specialized abilities which we had no intention of measuring. The obvious way to meet both these difficulties is to subject the data to a formal factor analysis.

In our own work we have followed the procedures recommended and used by Dr. Barakat and Dr. Moursy in their recent papers.²

(a) *With Bipolar Factors.* The matrix of observed correlations, given above in Table II, was first factorized by 'simple summation'. Three successive approximations (with incidental adjustments) sufficed to secure appropriate values for the reduced self-correlations. With an 8×8 matrix, the observed coefficients will provide 28 degrees of freedom; four summational factors would account for all but two. Four were accordingly extracted; but the last was too small to be statistically significant, and seemed devoid of any assignable meaning. The saturations for the three largest are shown in Table V.

(I) The first is a general factor, contributing over 60 per cent. to the total variance. It would presumably be identified, by those who prefer a summational procedure, with the general ability which the tests were designed to measure. The correlation between the factor-measurements for the individual testees and the independent assessments for intelligence provided by the teachers amounted to 0.67. This is certainly an improvement on the value obtained by taking an unweighted sum of the marks for the several tests (0.59).

(II) The second factor contributes more than 10 per cent. to the total variance; and was quite unexpected in its nature. It classifies each battery into two distinct parts. We at first supposed this might result from differences in the material used for the four subtests. For the first two it consisted of coloured blocks or insets to be manipulated by the child; for the second two, of black-and-white patterns on paper. But, after securing introspections from some of the children, and trying similar tests with students, we found that the material in which the problems were embodied had far less influence than the type of activity they involved: in the first two subtests the dominant process was apparently perceptual combination or construction; in the other two it was perceptual discrimination: in short, the first two seemed to depend largely on mental synthesis, the latter on mental analysis.³

(III) The third factor reveals an obvious contrast between the results of the first and second trials respectively. With this method of analysis it contributes rather less to the total variance than might have been anticipated, namely, 5.5 per cent.

In the light of Burt's comparison between the results of an ordinary analysis of variance and those of a summational analysis (this *Journal*, I, 12 and 25, Tables VI and XV), we were originally tempted to identify the three foregoing factors with the three components which nearly always underlie the analysis of variance in cases such as ours, namely, those representing (1) differences between 'Persons',

¹ A somewhat similar assumption is adopted by Kuder and Richardson ('The Theory of the Estimation of Test Reliability', *Psychometrika*, II, 1937, pp. 151-60).

² M. K. Barakat, 'Factorial Study of Mathematical Ability', this *Journal*, IV, 1951, pp. 144 f. (who gives detailed references for the requisite computational procedures), and E. M. Moursy, *ibid.*, V, 1952, pp. 166 f.

³ The relative superiority of the student or pupil in one or other of these forms of test often seemed to correspond with his intellectual type in other kinds of work. The distinction between analytic and synthetic types is not entirely new. We subsequently learnt that, when using various forms of non-verbal test (picture-completion, analogies, matrices, etc.), Burt and Moore had also noted a somewhat similar contrast: (cf. 'Experimental Tests of Higher Mental Processes', *J. Exp. Ped.*, V, p. 125, and *Brit. J. Educ. Psychol.*, XIX, p. 192).

(2) 'Interactions between Persons and Tests', and (3) 'Interaction between Persons and Trials' (or 'Occasions', as Burt prefers to say : cf. Table I). But the results of the further analysis of variance described above, and still more of the group factor analysis which we subsequently carried out, made it clear that this identification was only partly correct and might easily prove misleading.¹

TABLE V. ANALYSIS BY SIMPLE SUMMATION :
GENERAL AND BIPOLAR FACTORS

Tests	I	II	III
First Trial			
1. Picture	·924	·251	·223
2. Block	·860	·371	·142
3. Maze	·721	—·430	·317
4. Matrix	·781	—·309	·213
Second Trial			
1. Picture	·877	·291	—·154
2. Block	·804	·345	—·211
3. Maze	·803	—·340	—·185
4. Matrix	·226	—·178	—·344
Factor Variance	5·015	·831	·437
Contribution to Total Variance (per cent.)	62·7	10·4	5·5

TABLE VI. GROUP FACTOR ANALYSIS : BASIC, GROUP,
AND SPECIFIC FACTORS

Tests	Basic	Anal.	Synth.	1st Day	2nd Day	Specifics			
First Trial									
1. Picture	·791	·452	·098	·341	—·045	·153			
2. Block	·689	·523	—·058	·337	·108		·312		
3. Maze	·597	—·076	·501	·412	—·064			·187	
4. Matrix	·608	·101	·464	·376	·013				·255
Second Trial									
1. Picture	·827	·392	—·033	·052	·290	·153			
2. Block	·753	·431	—·008	—·106	·276		·312		
3. Maze	·701	—·112	·422	·130	·381			·187	
4. Matrix	·724	·086	·387	—·076	·409				·255
Factor Variance	4·094	·857	·808	·578	·493	·047	·194	·070	·130
Contribution to Total Variance (per cent.)	51·2	10·7	10·1	7·3	6·1	0·6	2·4	0·9	1·6

(b) *Group Factor Analysis.* Following Barakat we decided to carry out rotations both with Thurstone's graphical method and with Burt's arithmetical method. We began with a blind graphical procedure, rotating the summation factors two at a time in accordance with Thurstone's instructions.²

¹ In a personal note Professor Burt tells me that he himself would *not* make such an identification. "In the comparison printed in this Journal", he writes, "I took an extremely simple example: two tests only were applied on each day; hence, in this case both interactions *could* be represented by a single row of 'factor measurements' (cf. Table III, *loc. cit.*), and these 'factor measurements' might be expected to evince a close resemblance to those calculated from the bipolar (summational) factor analysis. In general, however, when n subtests have been applied, the residual matrix containing the interaction between the persons and the subtests will be a $2n \times N$ matrix, and will involve $(n-1)$ degrees of freedom, not one. Thus, with four subtests, the matrix would have an order of $8 \times N$, and *three* degrees of freedom, not one. Hence, unless the sample was so small as to nullify any additional factors, a single bipolar factor could not possibly account for it." And it was on his recommendation that we proceeded to undertake the group factor analysis. See also footnote (2), p. 120 above.

² Cf. Barakat, *loc. cit.*, p. 145. Thurstone, *Multiple Factor Analysis*, pp. 319 f.

Test Reliability in Terms of Factor Theory

Without changing to 'oblique' or correlated factors, it seemed impossible to reach anything like a perfect 'simple structure': there were, however, several alternative structures which would reduce the variance of the general factor to about 20 or 25 per cent. of the total variance. With the 'quadrimax method' it was reduced to 27 per cent. The reliability coefficients reached by such methods differed widely, and there seemed no definite reason for preferring one estimate to the other. At the same time there appeared to be strong grounds for rejecting all of them. Their multiplicity itself seemed to show that they were at once arbitrary and ambiguous. And above all, since the very conception of reliability implies the presence of a factor common to the various trials and the various tests, it seemed evident that procedures designed to abolish or diminish the general factor were quite unsuited to our problem.

We then attempted an arithmetical rotation, in accordance with the formulae and working methods described in this *Journal* (III, pp. 53 f.). We started by extracting one basic and four non-overlapping group factors. The lines of division between the groups were, as usual, determined by the arrangement of positive and negative saturations in the previous bipolar factors. After extracting the five factors corresponding to the significant factors in the preliminary summational analysis, we found fairly large residuals still remaining in the principal diagonal of the N.E. quadrant. These plainly indicate the presence of what would ordinarily be called 'specific factors' peculiar to each of the subtests. The repetition of the trials converts these specific factors into narrow group-factors or 'doublets'.

Accordingly, to secure a better approximation, a fresh analysis was carried out after these quasi-specifics had been eliminated. For this purpose the self-correlations in the N.E. quadrant were first 'reduced' by deducting the effects of the specifics, and five group factors extracted as before. To obtain saturations for the overlapping group factors a rotation-matrix was then calculated by the usual formula (*loc. cit.*, pp. 61 f.), and applied to the summational factor matrix. The results finally reached are set out in Table VI.

As usual, it was found that the 'basic factor' accounts for rather less of the variance than the corresponding 'general factor' obtained by the summational procedure. On the other hand, it correlates more highly with the teachers' assessments for general intelligence (0.76 instead of 0.67), and manifestly yields a better estimate of the ability our battery was designed to measure. The next two factors apparently represent specialized abilities for what we have tentatively called 'mental analysis' and 'mental synthesis' respectively—two irrelevant abilities which our battery was not intended to measure. The last pair of group factors represent the special conditions peculiar to one or the other of the two occasions on which the children were tested (including no doubt minor unintentional differences between the two forms of the battery).

Comparison of the Analyses. As we have seen, if we may accept the teachers' assessments of the pupils' general intelligence as a fair overall criterion, the weighted factor measurements for the 'general factor' provided by the summational analysis would possess a higher validity than the unweighted means of the test-scores (to which the analysis of variance refers), and those for the 'basic factor' provided by the group factor analysis a validity that is higher still.

With both methods of factorization there is clear evidence for what we could not have discovered from the analysis of variance, namely *the presence of specialized abilities for two contrasted types of test, 'analytic' and 'synthetic'*. These specialized abilities and the effects of the two different trials seem to be represented much more clearly and more naturally by pairs of group factors than by single bipolar factors. The group factor analysis moreover discloses what the summational or bipolar analysis could scarcely be expected to reveal, namely, *the presence of quasi-specifics each peculiar to the two different forms of the same subtest*. Now that these various supplementary factors have been detected by factor analysis, we could, if we wished, go on to test their significance by a more elaborate analysis of variance (here they are fully significant); and accordingly it seems plain that in future investigations the factor analysis might well be undertaken at the very outset.

In the present inquiry, with the 'factorial analysis of variance' the component for persons contributes 58 per cent. to the total variance. With the summational analysis the general factor contributes nearly 63 per cent. With the group factor analysis the basic factor contributes only 51 per cent. But, since the basic factor corresponds more closely with the ability we set out to measure, the last figure, which is also the lowest, here provides the safest guide.¹ Apparently with the analysis of

¹ Of course, had our aim been to test the combined resultant of *all* these abilities, this conclusion would no longer hold good. The analysis thus makes it clear that, as Burt puts it, "our formula for reliability must depend in part upon our specification of what the test is intended to measure. In that sense, the reliability we aim at is the reliability of the test-measurements, not merely of the test material. . . . Thus, if our aim is to measure 'innate general ability', which by definition does not change or fluctuate, we may, with the same test-problems, get at least two alternative sets of measurements, namely, measurements expressed as multiples of the standard deviations of the age-group and measurements expressed as I.Q.'s. The latter nearly always have a lower reliability. Weighting the tests will introduce further differences. For this reason, as already emphasized, the word 'test' is to be regarded as a shorthand term covering the whole method of measurement." These comments deserve re-quoting, because the points they bring out have apparently been overlooked by several recent critics (e.g., by A. Heim in her chapter on 'Test reliability' in *The Appraisal of Intelligence* (1954), pp. 67 f.).

variance and with the summational analysis the results obtained were swollen by the inclusion of some of the variance attributable to the more specialized abilities.

With the analysis of variance the proportion attributed to the interaction between persons and tests was 19 per cent. This value ($\sigma_{pt}^2 = \bar{r}_{sd} - \bar{r}_d$), it will be remembered, was presumed to represent the influence of specific factors tending to raise the 'self-correlations' for two applications above the 'reduced value' that we should expect if the irrelevant effects of these 'specifics' were eliminated. According to the group factor analysis, however, this estimate was excessive: for the four specifics lumped together it should apparently be nearer 5.5 per cent. The exaggerated estimate was doubtless due to the fact that, with the analysis of variance, the influence of the two group factors for mental synthesis and mental analysis (which was not separated out in the analysis of variance) has, in part at any rate, got incorporated in the interaction between persons and tests. Here again, therefore, the group factor analysis and the use of appropriate weighting seem more trustworthy.

The same holds true of the estimates for the influence of the different trials. For our present purposes this component is the most important of all, since it indicates the relative instability of the measurements obtained from virtually the same battery on different occasions. According to the analysis of variance it accounted for 11 per cent. of the total variance; according to the bipolar analysis, for only 5.5 per cent.; and according to the group factor analysis it contributes 7 per cent. on the first occasion and 6 per cent. on the second, i.e., 13 per cent. in all. In the present experiment the peculiar conditions distinguishing the two different trials are most naturally represented by two different group factors: to convert these two group factors into a single bipolar factor (as is done explicitly by the summational analysis and implicitly by the analysis of variance) may easily distort the facts and underestimate the influence of such disturbing conditions. All these arguments, therefore, point to the same conclusion, namely, that, for such problems as the present, *the factorial analyses are superior to the analysis of variance*, and of the two factorial procedures *the group factor method is superior to the bipolar or summational*.

Final Assessment of Reliability. From the foregoing results one practical corollary stands out in sharp relief: before the investigator can estimate the reliability of his test, or even decide on an adequate procedure for so doing, he must first settle the question—reliability for what? Here the answer is quite plain: our tests were constructed to measure 'general intelligence'; and, from the evidence available, there can be little doubt that our best approximation to this is to be found in the 'basic factor' common to all the tests.

To obtain the best assessments for this basic factor, the obvious procedure is to compute the regression coefficients by the usual formula $f'R^{-1}$, where f' is the row-vector of saturations for the basic factor (Table VIII, col. 1) and R^{-1} the inverse of the relevant correlation matrix (Table II). Then, instead of taking the unweighted sum of each pupil's marks for each subtest (p. 124) we shall use the regression coefficients to weight the marks obtained at the first and the second trial respectively.¹

To obtain an improved estimate for reliability based on the factor measurements, the requisite equation is similar in form to that already given (eqn. ii); but the row vector w' will now consist of differential weights (the regression coefficients) instead of equal weights. The reliability coefficient thus computed proves to be 0.838. This figure therefore provides the most appropriate estimate of reliability for our present purpose.²

We should, however, like to insist that the determination of a single final figure for reliability, however ingeniously reached, is by no means the sole or the main conclusion to be derived from such analyses as the present. As the reader will have perceived, the alternative and incidental calculations described in the foregoing sections have given that figure a clearer meaning, and at the same time have greatly increased our detailed understanding of the tests and of the points at which they need to be revised. Here, however, our object was not to report any specific results achieved by this particular inquiry, but merely to use our data to illustrate what we consider to be improvements in current statistical techniques.

¹ These regressions are themselves nearly always instructive. Here, for example, small negative weights are found for the Block and Matrix tests, which thus operate as 'suppressor-variables', and counteract the effects of the two irrelevant specialized abilities. The Picture test has the highest weight, and would seem to indicate the most reliable type of subtest for a population of this kind. We are therefore hoping to improve the efficiency of the other tests by reconstructing them with more colourful and pictorial material.

² It might have been expected that the introduction of negative weights would reduce the value now obtained for the reliability coefficient below that previously reached with equalized weights. With equalized weights, the figure for reliability would (one might suppose) be increased by the inclusion of the more specialized abilities. That influence is certainly traceable; but its removal seems to be more than counterbalanced by other improvements in the estimate which the differential weighting has produced. The subtests which have an intrinsically poor reliability (the Matrix test for example) now receive a very low weight. In other investigations my co-workers and I have frequently found that the use of negative weights and the elimination of irrelevant abilities *do* reduce the reliability coefficient, though they nearly always increase the validity coefficient (i.e., the multiple correlation with the general or basic factor). At times therefore—indeed, it would seem, more often than not—the use of the simpler formula for the test-retest coefficient may decidedly exaggerate the true reliability; and for this reason we believe that, in all cases of doubt, it is essential to use a group factor analysis to estimate reliability as well as validity.

INDEX TO VOLUME VIII

INDEX OF AUTHORS

<i>Authors</i>	<i>Titles</i>	<i>Page</i>
Burt, C.	Sir Godfrey Thomson	1-2
Burt, C.	Test Reliability Estimated by Analysis of Variance	103-118
Burt, C., Cooper, W. F., and Martin, J. L.	A Psychological Study of Typography	29-57
Cattell, R. B., and Cattell, A. K. S.	Factor Rotation for Proportional Profiles	83-92
Cureton, E. E.	On the Use of Burt's Formula for Estimating Factor Significance	28
Guttman, L.	An Additive Metric from all the Principal Components of a Perfect Scale	17-29
Guttman, L.	The Determinacy of Factor Score Matrices with Implications for Five Other Basic Problems of Common-Factor Theory	65-81
Mahmoud, A. F.	Test Reliability in Terms of Factor Theory	119-135
Roberts, J. A. F., and Dunsdon, M. I.	A Study of the Performance of 2,000 Children on Four Vocabulary Tests	3-16
Simon, H. A., and Guetzkow, H.	Mechanisms Involved in Group Pressures on Deviate- Members	93-101
Stuart, A.	The Correlation between Variate Values and Ranks	25-27

INDEX OF SUBJECTS

<i>Titles</i>	<i>Authors</i>	<i>Page</i>
Factor Score Matrices, the Determinacy of, with Implications for Five Other Basic Problems of Common-Factor Theory	Guttman, L.	65-81
Factor Significance, On the Use of Burt's Formula for Estimating	Cureton, E. E.	28
Group Pressures on Deviate Members, Mechanisms involved in	Simon, H. A. and Guetzkow, H.	93-101
Principal Components of a Perfect Scale, An Additive Metric from all the	Guttman, L.	17-29
Proportional Profiles, Factor Rotation for	Cattell, R. B., and Cattell, A. K. S.	83-92
Ranks, The Correlation between Variate Values and	Stuart, A.	25-27
Test Reliability Estimated by Analysis of Variance	Burt, C.	103-118
Test Reliability in Terms of Factor Theory	Mahmoud, A. F.	119-135
Thomson, Sir Godfrey	Burt, C.	1-2
Typography, A Psychological Study of	Burt, C., Cooper, W. F., and Martin, J. L.	29-57
Vocabulary Tests, A Study of the Performance of 2,000 Children on Four	Roberts, J. A. F. and Dunsdon, M. I.	3-16

BOOKS REVIEWED

<i>Author</i>	<i>Title</i>	<i>Page</i>
Bowden, B. V. (ed.)	Faster than Thought: A Symposium on Digital Computing Machines	59-64

OCCUPATIONAL PSYCHOLOGY

Editor : ALEC RODGER

January, 1955

CONTENTS

Volume 29, No. 1

A Method of Interviewing used in Studies of Workers' Attitudes :

1. Effectiveness of the Questions and of Interviewer Control.

By R. MARRIOTT and R. A. DENERLEY

Comparison of Paced and Unpaced Performance at a Packing Task.

By R. CONRAD, assisted by BARBARA A. HILLE

The Efficiency of Labour in Coalmining

By W. H. SALES

The Ishihara Test and Defects of Colour Vision.

By PETER CAVANAGH

Book Reviews

Other Books Received

Obituary

Annual Subscription 30 shillings

National Institute of Industrial Psychology
14, Welbeck Street, London, W.1

HUMAN RELATIONS

A Quarterly Journal of Studies towards the Integration of the Social Sciences

Contents of VOL. VIII, No. 1

F. KRÄUPL TAYLOR. The Three-Dimensional Basis of Emotional Interactions in Small Groups. II.

JOHN T. LANZETTA. Group Behavior under Stress.

CECILY DE MONCHAUX and SYLVIA SHIMMIN. Some Problems of Method in Experimental Group Psychology.

ROBERT R. BLAKE, MILTON ROSENBAUM, and RICHARD A. DURYEA. Gift Giving as a Function of Group Standards.

DAVID P. AUSUBEL. Sociempathy as a Function of Sociometric Status in an Adolescent Group.

Book Reviews.

30s. per annum or 8s. 6d. per issue

The Tavistock Institute of Human Relations
London, England.

The Research Center for Group Dynamics
Ann Arbor, Mich.

TAVISTOCK PUBLICATIONS LTD

2 Beaumont Street, London, W.1

FW

CONTENTS

		<i>Page</i>
OBITUARY	Sir Godfrey Thomson	1
ROBERTS, J. A., AND DUNSDON M. I.	A Study of the Performance of 2,000 Children on Four Vocabulary Tests	3
GUTTMAN, L.	An Additive Metric from all the Principal Com- ponents of a Perfect Scale	17
STUART, A.	The Correlation between Variate Values and Ranks	25
CURETON, E. E.	On the use of Burt's Formula for Estimating Factor Significance	28
BURT, C., COOPER, W. F., AND MARTIN, J. L.	A Psychological Study of Typography	29
NOTE	The International Conference on Factor Analy- sis : Programme of Papers	58
BOOK REVIEW	Digital Computing Machines	59

JOURNAL COMMITTEE

CYRIL BURT *Editor*CHARLOTTE BANKS *and* ALAN STUART *Assistant Editors*J. W. WHITFIELD *Managing Sub-editor**and*

B. M. FOSS

J. S. SMALL

M. HAMILTON

J. SUTHERLAND

A. HERON

F. W. WARBURTON

